DOCUMENT RESUME

ED 170 339 TM 008 654

AUTHOR Maling-Keepes, Jillian
TITLE Educational Evaluation: Key Characteristics. ACER Research Series No. 102.
INSTITUTION Australian Council for Educational Research, Hawthorn.
PUB DATE 76
NOTE 161p.; Figures may not reproduce well due to small print size

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Curriculum Evaluation; Data Analysis; Data Collection; *Educational Assessment; Educational Objectives; *Educational Programs; *Evaluation Methods; *Evaluators; Formative Evaluation; Information Utilization; Program Development; *Program Evaluation; Research Design; Summative Evaluation

ABSTRACT

A set of 13 key characteristics is presented as a framework for educational evaluation studies: (1) program's stage of development when evaluator is appointed; (2) program's openness to revision; (3) program uniformity from site to site; (4) specificity of program objectives; (5) evaluator's independence; (6) evaluator's orientation to value judgments on program merits; (7) evaluator's orientation to utilization of his work; (8) attention to program goals; (9) generalizability; (10) formality of control; (11) emphasis on comparison; (12) scope of information collected; and (13) emphasis on interpretation. These characteristics are divided into four groups related to: the program being evaluated; the evaluator's position; questions asked in the evaluation; and evaluation design, and collection and interpretation of information. Each characteristic is further defined by more specific aspects of the evaluation or the program. Eight existing evaluation studies illustrate and compare the application of these characteristics: Neighborhood Benefit Program; Psychiatric Inpatients; Performance Contracting Experiment; Experimental School, Nuffield A-Level Biological Science Project; Mathematics Project; Humanities Curriculum Project; and Educational Extension Agents. An extensive bibliography is appended. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED170339

TM008 654

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

EDUCATIONAL EVALUATION
key characteristics

Jillian Maling-Keepes

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Australian Council for Educational Research

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Educational Evaluation:

Key Characteristics

ACER Research Series No. 102

Educational Evaluation:

Key Characteristics

Jillian Maling-Keepes

Australian Council for
Educational Research

Published by
The Australian Council for Educational Research, Limited,
Frederick Street, Hawthorn, Victoria 3122
December 1978

Typeset by
Savage & Co. Pty Ltd,
Fortitude Valley, Qld 4006

Printed and bound by
Brown Prior Anderson Pty Ltd,
Burwood, Victoria 3125

National Library of Australia Cataloguing-in-Publication data.

Maling-Keepes, Jillian Mary
Educational evaluation: key characteristics.
(Australian Council for Educational Research series; no. 102).
Bibliography.
ISBN 0 85563 180 5.

1. Education – Evaluation. I. Title. (Series).

370

Copyright © Jillian Maling-Keepes 1976
No part of this publication may be reproduced in any form without permission from the publisher.

For Lee, Decker, and Bruce

Contents

	Page
1: Introduction	1
Statement of the Problem	1
Uses and Implications of the Set of Characteristics	3
2: Background	4
Evaluation Prior to 1960	4
Increased Demand for Evaluation in the 1960s	8
Response to the Demand for Evaluation	10
Recent Concepts of Evaluation	11
Difficulties with Recent Theory Building	17
Efforts to Clarify the Confusion	21
3: Development of the Characteristics	24
Purpose of the Characteristics	24
Form of the Characteristics	25
Criteria Used in Developing the Characteristics	26
Relationship of the Characteristics to Theory	28
4: Presentation of the Characteristics	31
Characteristics Related to the Program	32
Characteristics Related to the Evaluator's Stance	46
Characteristics Related to Questions Asked in the Evaluation	57
Characteristics Related to Evaluation Design, Information Collection, and Interpretation	63
5: Examples of Use of the Characteristics to Describe Evaluations	78
6: Postscript	124
Summary	124
Conclusions	128

Page

APPENDIX: THE CHARACTERISTICS AND THEIR ELEMENTS 131

BIBLIOGRAPHY 140

FIGURES

1 Summary of the Characteristics 31
2 Summary of the Case Studies 80
3 Characteristic I: Program's Stage of Development when the Evaluator Is Appointed 83
4 Characteristic II: Program's Openness to Revision 86
5 Characteristic III: Program Uniformity from Site to Site 89
6 Characteristic IV: Specificity of Program Objectives 92
7 Characteristic V: Evaluator's Independence 96
8 Characteristic VI: Evaluator's Orientation to Value Judgments on Program Merits 100
9 Characteristic VII: Evaluator's Orientation to Utilization of his Work 103
10 Characteristic VIII: Attention to Program Goals 106
11 Characteristic IX: Generalizability 109
12 Characteristic X: Formality of Control 112
13 Characteristic XI: Emphasis on Comparison 115
14 Characteristic XII: Scope of Information Collected 119
15 Characteristic XIII: Emphasis on Interpretation 123

1

Introduction

STATEMENT OF THE PROBLEM

Prior to 1960, evaluation – apart from 'accreditation' – was generally equated with measurement and, even more specifically, with testing: the characteristics of the one were those of the other. In the early 1960s this changed. There was a sharp increase in the number of evaluations being done, and those conducting the evaluations came from diverse backgrounds – different disciplines (psychology, sociology, philosophy, etc.) – and brought different expertise (measurement, research design, participant observation, and occasionally curriculum). They worked on a wide variety of projects at different stages of development and dissemination. The 60s saw widespread use of criterion-referenced tests and matrix sampling. More extensive use was made of questionnaires, interviews, classroom observation schedules, cost-effectiveness analysis, self-reports, diaries, case studies, and ethnographic techniques. The new techniques in part indicate the diversity of skills of evaluators, and in part the diversity of projects.

The theorizing about the nature of evaluation reflects the changes in practice already described, a rapid increase in volume and variety of evaluation, and little systematic attention paid to it.

> The accelerating development of research in the area of educational evaluation has created a collection of concepts, facts, generalizations, and research instruments and methods, that represent many inconsistencies and contradictions because new problems, new conditions, and new assumptions are introduced without reviewing the changes they create in the relevance and the logic of the older structure. (Tyler, 1967, p.13)

Something of the inconsistencies and contradictions Tyler refers to can be seen in the definitions and terms used for evaluation. At present these exist side by side, which suggests widely different practices and cloaks commonalities. There are competing conceptualizations and diverse practices, and no framework to delineate the features which,

while common to evaluation studies, are significant in differentiating them.

At the same time, there are demands that evaluation studies themselves be evaluated. Although some wish to establish standards for educational evaluations, little attention has been paid either to the nature of the phenomena or to establishing prior consensus as to evaluation's key aspects. Curriculum workers, such as Westbury (1970) and Eisner (1972), have been particularly critical of educational evaluation and, without a comprehensive framework, have been unable to articulate fully whether their argument is with evaluation per se or particular forms of it. Yet there is no view of evaluation which, while delineating its key aspects, gives an adequate account of its diversity and complexity.

It is the purpose of this study to describe a framework for evaluation, drawing on the many suggestions made in the conceptualizations and accounts of evaluations by evaluators in and outside of education during the period 1960–1974. The framework is presented as a set of key characteristics of educational evaluation.

The characteristics are abstractions, related to evaluation, that serve to discriminate between one evaluation study or conceptualization and another. There are two parts to each characteristic: the characteristic is named and its elements identified. The characteristic's name refers to a broad abstraction (such as Formality of Control), whereas the elements indicate more specific aspects of evaluations or the programs evaluated. The elements of a characteristic are closely inter-related, which indicates that there are factors which need to be considered in relation to each other in arriving at a description of the evaluation in terms of the broad abstraction. For example Characteristic VII, Utilization, includes as elements: identification of those likely to use the evaluation; uses to which the information is likely to be put; and the form in which the information is presented. An evaluator may identify those likely to use his study, yet not consider possible users, and present information in an unintelligible format. In effect, the evaluator's position is half hearted and hence possibly ineffective in this respect. In characterizing the evaluator's orientation to use, the three elements taken together indicate a different picture of the evaluation to any one element taken singly. The characteristics included are 'key' ones. They are abstractions that make a difference in an evaluation, that is in the way it is conducted and in its subsequent interpretation or standing. For example a characteristic on 'Evaluator's Independence' has been included in the final list. A characteristic on the 'texture of information' collected has not been included. The evaluator's independence affects the way others will subsequently

treat the evaluation study, and it also relates to the conduct of the study itself. The texture of information collected in the study, whether closely woven or loose, is an interesting feature of an evaluation but does not tell us anything about the study beyond that particular detail. Characteristics that excluded evaluations of a particular kind, as in the case of 'Data Analysis', have also been omitted. Similarly items associated with evaluations, but not forming part of them, have been set aside. For example the publication policies of academic journals may well influence aspects of evaluations and the style used in such studies, but they do not form part of the evaluation itself or part of the evaluator's design.

Another person would conceive of the characteristics of educational evaluations in different terms to those used here. The labels used for characteristics would differ as would the elements listed. In this sense, the set of characteristics listed is not definitive, yet, if the set is reasonably comprehensive, a second listing of educational evaluation characteristics could be mapped on to those given here. Similarly some gauge of the comprehensiveness of the characteristics included here is given by the extent to which they afford a basis for comparing concepts of evaluation advanced by different writers. Such a comparison is sketched in Chapter 6.

USES AND IMPLICATIONS OF THE SET OF CHARACTERISTICS

As a set, the characteristics afford a basis for developing a structured description of an evaluation study or conceptualization. They thus facilitate comparisons, several of which are useful to those conceptualizing evaluation and to those conducting evaluation studies. They may also function as an aid to someone designing an evaluation study and may act as heuristics, indicating possible lines of inquiry into the work of evaluators. These possible uses too are detailed in Chapter 6.

2

Background

Evaluation prior to 1960

Prior to the advent of the national curriculum projects in the late 1950s there were, perhaps, three evaluative approaches in education: the use of expert judgments, measurement, and Tyler's emphasis on objectives, (Guba, 1969; Gephart, 1972; Worthen & Sanders, 1973). The last two of these were closely linked with testing – so closely that some do not distinguish between them.

The use of expert judgments in education is well exemplified by the development of the accreditation agencies in the 20s and 30s (Gephart, 1972). This development called for assembling a team of experts who made a site visit, observed the institution operating, and met with officials and others affected by the institution, and probably used some kind of check list to guide their judgment. The visiting team deliberated on their work and compiled a report on the institution in question. Their report presented their verdict on the institution and its program (Worthen & Sanders, 1973). In this form of evaluation, which continues today, emphasis is on the personal expertise of those making the judgment. Experts can take into account the interplay of many factors, in a relatively brief period, with little time lag developing between the collection of information and the final evaluative judgment. The expert-judgment process, however, relies heavily on what those interviewed are willing to tell – and, given that the experts come from 'outside' and usually from 'higher up', there is more than a possibility that the information will not be candid and could even be heavily guarded (Weiss, 1972b). The chief reliance and both strength and danger of this procedure lie in the group of experts themselves. If they are perceptive, skilful, insightful, and impartial in outlook, their evaluation can be brilliant; if they are not, it is quite likely to be seriously wide of the mark, as superficial as television talk shows. Further, since the process by which the judgment is arrived at remains largely implicit, there have been fears as to the

reliability and objectivity of the procedures (Guba, 1969). As a result, many recent writers speak of evaluation in terms which exclude accreditation or the judgmental approach.

The second major approach in evaluation prior to 1960 emphasized measurement. Central was the thesis:

> Whatever exists at all exists in some amount ... Anything that exists in amount can be measured ... Measurement in education is in general the same as measurement in the physical sciences. (McCall, 1922, pp.1–5)

Those associated with the measurement approach developed a host of measuring devices. A good deal of stress was laid on tests involving recall of factual information as indicative of the ability to reason or think more widely about the relevant subject matter (Buckingham, 1916; Thorndike, 1914; Wood, 1923). Thorndike also argued,

> We want all sorts of scales; wherever there is an ability, or skill, a product that needs to be measured, there we need a scale. Of course, the task is endless (Thorndike, 1914, p.147).

McCall asserted that only the lack of ingenuity of scientists would prevent measurement of 'the love of a sunset or appreciation of an opera' (McCall, 1922, p.6). The measurement movement readily adopted Spearman's discovery of the effect of random error on test scores and his use of the product-moment coefficient for predicting the reliability of tests of varying lengths. Validity received similarly widespread, although perhaps less successful, attention. Thus the measurement approach generally favoured operations that could be handled statistically. The link between measurement and evaluation continues today. Its closeness can be seen in the many texts titled 'Measurement and evaluation'. (See, for example, Thorndike and Hagen (1961), *Measurement and evaluation* and more recently, Karmel (1970), *Measurement and evaluation in the school*.) There are even instances where the terms measurement and evaluation are so closely linked as to be treated as a singular noun: '... measurement and evaluation is an integral part of the teaching process' (Lien, 1967).

The third major emphasis in evaluation prior to 1960 stemmed from Tyler's work, with its stress on congruence between the objectives and performance. In 1930 Tyler spoke of testing, not evaluation. An early article, 'Measuring the ability to infer' (Tyler, 1930), sets out two steps for test construction: 'inventing ways of measuring important aspects' of knowledge (in this case, zoology) and 'elimination of unreliability and subjectivity' (two of the preoccupations of the measurement movement). By 1931 the two steps of the earlier model had been expanded

to ten, Tyler's familiar emphasis on objectives emerged, while the emphasis on objectivity and reliability was kept:

1 Formulation of course objectives;
2 Definition of objectives in terms of student behaviour;
3 Collections of situations in which students will reveal presence or absence of each objective;
4 Presentation of situations to students;
5 Evaluation of student reactions in light of each objective;
6 Determination of objectivity of evaluation;
7 Improvement of objectivity, when necessary;
8 Determination of reliability;
9 Improvement of reliability, when necessary;
10 Development of more practicable methods of measurement when necessary.

In pointing out how this approach differed from conventional measurement practice, Tyler emphasized the centrality of objectives. His method proceeded from an analysis of objectives, not of content; test situations were selected from an analysis of the objectives by spelling out the behaviour of students, not by an arbitrary decision between item forms; validation was against objectives not against other tests, or teachers' marks.

By 1934, Tyler had adopted the term 'evaluation' for a generalized procedure for assessing human behaviour. His approach had twin central constructs: objectives and student behaviour. Both were defined broadly so that, for example, objectives did not refer just to specific elements of knowledge, but were defined as important outcomes of education, cognitive and affective, serving both teaching and evaluation.

The Eight Year Study (1934–1942) demonstrated the centrality of these two constructs while making apparent the reliance of the approach on a broad array of measures. Use was made of many different tests, scales, inventories, check lists, questionnaires, to gather information about how well a wide range of curricular objectives was being achieved in a group of schools (Smith & Tyler, 1942). The study also made clear the extent to which ideas on testing in 'Generalized techniques for constructing achievement tests' (Tyler, 1931) had become ideas on evaluation. There were only seven steps in the Eight Year Study for evaluation of educational outcomes. A comparison of these with the ten steps of 1931 shows that the major change lies in the alteration of the earlier steps that specifically dealt with reliability and objectivity to a broader notion of utility (Echols, 1973; Finn, 1972). Tyler's approach continues in the ideas about evaluation put forward by Provus (1969, 1972) and Metfessel and Michael (1967),

among others, although they define objectives and behaviour in a much narrower sense.

There was a lapse of almost 20 years between the Eight Year Study and the evaluation of curriculum projects at the beginning of the 60s. During that time, the hectic concentration on measurement had waned; but evaluation continued to be used very substantially in the context of improvement of instruction in individual classrooms. (See, for example, Dressel and Mayhew's work: Dressel & Mayhew, 1954 a; b; Dressel & Nelson, 1956; Dressel, 1961.)

It is likely that, with time, the measurement approach and that of Tyler have accommodated each other, and become fused. There always were similarities. Both revolved around some form of testing, although Tyler did advocate the use of other instruments; each stressed the scientific over the intuitive, the rational over the impressionistic; the Eight Year Study paid as careful attention to reliability and validity as did the measurers. Fusion of measurement and Tyler's approach into a single concept of evaluation may well have been hastened by the fact that few – Wrightstone's work is the outstanding exception – subsequent evaluation studies fully followed the paradigm embodied in the Eight Year Study (Finn, 1972). The complexity of procedures, instrumentation, and interpretation used in the Eight Year Study may have been simplified, between 1940 and 1960, to the much more manageable one of pre- and post-achievement tests, while still retaining Tyler's stress on objectives. For example, Hastings described evaluation prior to the curriculum projects of the 60s in these terms:

> For many years the expression 'curriculum evaluation' has tended to mean to most people some sort of use and interpretation of achievement tests ... the most commonly held idea of the sequence of evaluation endeavours starts with the art of stating the objectives of a set of materials ... This is followed by definition of the objectives in behavior terms ... Next comes the development of items ... These items are combined into scorable units ... Then finally the sequence ends in attempts to interpret these scores. (Hastings, 1966, p.27)

The emphasis on objectives expressed in behavioural terms is as Tylerian as the description of the process. Yet testing, typical of the measurement approach, is the central procedure.

The concept of evaluation described by Hastings was not limited to curriculum evaluation, nor indeed to education. Weiss, speaking of action programs, defined the 'traditional' approach to evaluation as:

> Finding out the goals of the program; translating the goals into measurable indicators; collecting data on the indicators for those who've been exposed to the program; collecting similar data on a control group; com-

paring the two sets of data in terms of the goals of the program. (Weiss, 1972a, p.6)

If 'achievement tests' was substituted for 'measurable indicators', the only element present in Weiss's account and missing from Hastings's would be a control group.

INCREASED DEMAND FOR EVALUATION IN THE 1960s

The first wave of 'Great Society' legislation in the United States, whether new or extending existing programs, affected many areas of health, welfare, the legal system, and education. These programs, aimed at improving some aspect of the human condition, involved huge investments of money, time, and energy. It was not surprising then that questions as to the effectiveness of these programs became commonplace (Suchman, 1967, p.1; Worthen & Sanders, 1973, p.5; Caro, 1971a, pp.ix-x). How well did the many programs, new and old, succeed? Which succeeded? Under what conditions? Ways had to be found for determining which of them was successful so that past and future expenditures could be justified. The demand for systematic feedback gave impetus to the development of evaluation in the early 60s.

In education, evaluation received major impetus from at least three sources: Sputnik, the Elementary and Secondary Act of 1965, and, most recently, from demands for accountability. The launching of Sputnik stimulated demands for curriculum reform. From the late 50s through the 60s there was massive funding of curriculum projects: the School Mathematics Study Group, the Madison Project, the Biological Sciences Curriculum Study program, and Harvard Project Physics, among others. Those supporting the development of new curricula, as well as administrators, teachers, and communities who would use them, requested information on the courses' effectiveness, frequently in comparison with other courses in the same area. By 1966, Forehand was able to assert:

> In almost every formal curriculum development project there is a person responsible for the measurement and evaluation of the curriculum's effect upon student performance. (Forehand, 1966, p.199)

The development of large-scale curriculum projects in England, although slightly later, parallels that of the USA. Evaluation was integral to the Nuffield curriculum projects. Since 1965, extensive curriculum development work undertaken by the Schools Council has been accompanied by some form of evaluation (Entwistle & Nisbet, 1972; Schools Council, 1973).

Australia, too, saw developments in the evaluation of curricula. Some of these were associated with large-scale projects of the kind represented by the Australian Science Education Project. On the whole, the large majority of evaluation studies in Australia were undertaken by the curriculum and research branches in the various States and Territories – and indeed this was one of the reasons for the rapid development of such centres during the decade between 1965 and 1975. This situation, in turn, meant that many of the studies and reports themselves were directed to an internal audience within the particular education department. Such reports filter more slowly through to the general education audience, which makes the task of charting the pattern of growth and development less open to public view. Further there are signs that Australia in the latter half of the 1970s was, perhaps, only just entering a major stage of curriculum development equivalent to that initiated in England with the Schools Council and in the USA in the early 1960s. The establishment of the Curriculum Development Centre in Canberra resulted in an increased number of curriculum projects being planned. Some have already been developed and disseminated. (The Social Science Education Materials Project is one such example.) As many are just starting, there are signs that the demand for evaluation is likely to increase in the next few years. It seems that many of the general background factors influencing the development of evaluation in the USA, England, and Australia are similar. While the USA is the example of the development discussed in this study, the general pattern of increased demand, diversity of task entailed in evaluation, and diversity of background of those understanding evaluation studies can be traced in each of the three settings: USA, UK, and Australia.

In the USA, demand for educational evaluation was greatly extended by the Elementary and Secondary Education Act (ESEA) of 1965 which contained the following proviso:

> Each educational agency shall, at least annually, provide an evaluation of the effectiveness of its Program ... in meeting the special education needs of the educationally deprived children, including appropriate objective measures of educational achievement. The measurement of educational achievement under such a program shall include the measuring or estimating of educational deprivation of these children who will participate in the program, and the comparing, at least annually, the educational achievement of participating children with some objective standard or norm. (Quoted in Law, 1971, p.41)

It is noteworthy that the legislation required that each project funded under Title I and Title III be evaluated, and that the evaluation was to focus on program impact on student learning and to be based on objective evidence, as contrasted with impressions and opinions.

The demand for evaluation made explicit in the ESEA legislation was part of a wider more diffuse demand for accountability. Occasionally this too surfaced in legislative action. California's Stull Bill (Senate Bill 90) mandated evaluation of teachers in terms of student performance. More commonly, accountability became a slogan of local and national groups, educators, and public alike. The National Advisory Commission on Civil Disorders asserted that, to increase the accountability of the public schools, the results of their performance should be made public. Similarly the Research and Policy Committee of the Committee for Economic Development argued:

> Innovation in education ... has become essential if the schools are to be genuinely effective in achieving their aims and goals. Continuing assessment of the product of the schools also is necessary. This means the development of principles and techniques for critically judging the worth of whatever the schools teach and the effectiveness and efficiency of their methods of instruction. (Committee for Economic Development, 1968, p.13)

Such statements became so prevalent that Lortie, a University of Chicago sociologist, predicted that a 'certificated public evaluator', serving a function analogous to the certified public accountant, would emerge.

RESPONSE TO THE DEMAND FOR EVALUATION

Not surprisingly, beginning in the early 1960s more attention was paid to evaluation. There was a marked rise in the number of seminars, workshops, symposia, and publications on evaluation. Professional associations turned their attention to evaluation. The American Educational Research Association (AERA) and the Association for Supervision and Curriculum Development (ASCD) put out special publications on evaluation. Journals brought out special issues on evaluation: *Urban Review*, 1969; *Journal of Research and Development in Education*, 1970; *Review of Educational Research*, 1970. New journals came into being: *Evaluation*; *Evaluation Comment*; *Evaluation Product Reports*; *Transaction*. Colleges and educational agencies established centres for the study of evaluation. One federally funded research and development centre was established solely for the study of evaluation. Programs for training evaluators developed under the impetus of the federally funded Research Training Program.

More important than the obvious signs of increased attention was the way in which the demand was met. At the beginning of the 60s, there was no cohort of trained evaluators waiting on the wing. Those who conducted evaluations did so in response to the immediate

demand. They came from different disciplines such as sociology (Weiss); psychology (Suchman, Louis Smith, Walberg); philosophy (Scriven); administration and organization (Stufflebeam). They brought to evaluation studies diverse interests and expertise: measurement, research design, participant observation, cost-benefit analysis, administrative and, occasionally, curriculum experience.

Evaluations were done under contract, within time limits, or limited budgets. Turnover of personnel on evaluation projects was often high. Frequently directors of evaluation projects worked in a part-time capacity. When they worked full-time on an evaluation project, it was likely to be a short-term commitment between more stable and long-term commitments of a permanent institutional relationship (Weiss, 1970).

What was to be evaluated was extremely diverse: as tangible as a box of materials, as nebulous as 'what teachers did', or as complexly unstructured as 'open classrooms' and 'alternative schools'. It could be paper and pencil instructional materials accompanied by film loops; but it was just as likely to be a theatre for underprivileged children, a mobile learning laboratory for migrant workers, teaching aids, or a computerized system for retrieving research information for teachers (Glass, 1968).

Projects varied widely in scale. Some were limited to a few classrooms within a single school district (typical of many Title I and Title III evaluations), others (e.g. the School Mathematics Study Group in the USA, the Humanities Curriculum Project in the UK) were tried out at selected sites across an entire country. Some of the programs of social action instituted in the mid 1960s (Project Headstart and Project Follow-Through, etc.) were on a new scale, aimed broadly at improving education nationwide, not at a few selected sites but for millions of children (Cohen, 1970, p.213).

Practices had to change. At one time, norm-referenced tests with pre-test/post-test only design were the mainstay of educational evaluation. Projects of the 60s called for increased use of criterion-referenced tests and new ways of using them (matrix sampling) – questionnaires, interviews, classroom observation schedules, role-playing, cost-effectiveness analysis, self-reports, diaries, case studies and ethnographic techniques (Merwin, 1969; Metfessel & Michael, 1967).

RECENT CONCEPTS OF EVALUATION

One index of the different views held on evaluation are the definitions of it introduced since 1960. These are dealt with here in some detail

in order to convey the sheer volume of concepts, their diversity, the differences between these concepts of evaluation and those held prior to 1960, and the associated problems of communication and confusion. A survey of terms and definitions does not take into consideration the many implied concepts of evaluation conveyed in reports and discussions that are not formalized in special language or definition. If these were included, the range of ideas would be further extended.

Some defined evaluation as a form of scientific research, and stressed the importance of experimental designs, control groups, statistical procedures, well-drawn samples, and so on. Thus Suchman chose to distinguish evaluation (i.e. making of a value judgment of any kind) from evaluative research on the grounds that in the latter scientific research techniques are used in making an evaluation. He defined evaluation as:

> those procedures for collecting and analysing data which increase the possibility for 'proving' rather than 'asserting' the worth of a social activity. (Suchman, 1967, pp.7-8)

Rossi (1969, 1972) concurred, and argued that research designs, statistical techniques, and data collection methods are the same for research into the most basic principles of human behaviour as for evaluation of a social action program. Indeed, many of those involved in evaluation of social action programs would include evaluation-as-research as part, if not all, of their definition of evaluation. (See, for example, Cain, 1974; Caro, 1971; Hyman & Wright, 1967; Weiss, 1970.)

The stress on evaluation as a form of research is shared by many educators. Payne repeated Suchman's definition almost verbatim (Payne, 1974, p.7). Harris, in 1963, defined evaluation as the 'systematic attempt to gather evidence regarding changes in student behaviour that accompany planned educational experiences' (Harris, 1963, p.191). In 1967, at a symposium on evaluation held at the University of California at Los Angeles, he repeated the definition, stressing that without application of research procedures the evaluator was unable to furnish 'real' evidence or proof. Others at that symposium stressed the value of evaluation-as-research not so much for proving the final judgment on a program, but in contributing to the estimate of cause-and-effect relationships (Wittrock & Wiley, 1970). More recently the view has received sustained expression in Cooley and Lohnes, *Evaluative inquiry in education* (1976). They asserted that evaluative inquiry shares the same rules of evidence, procedures, and standards as other forms of research, since evaluation is an applied science. The kind of research argued as appropriate is multivariate, large-sample, longitudinal, and, they admit, expensive. But, unlike Harris or Wittrock, they did define

evaluative research as furnishing information for decision-making – an emphasis that will be returned to later.

There are, then, many who see evaluation primarily as a form of research, yet they do not speak with a single voice – instead they present variations on a theme. Others in education who subscribe to some variant of the view include: Stanley, 1972; Campbell, 1969; Coleman, 1969; Cunningham, 1973; Larkins and Shavelson, 1969; Light and Smith, 1970; and Welch and Walberg, 1968. Aspects of the measurement approach live on, supported by increased data processing power of computers and new statistical techniques. Others have retained Tyler's emphasis on objectives, defining evaluation as the comparison of performance against standards drawn from the goals of the program, preferably stated in behavioural form:

> Evaluation is the process of (a) agreeing upon program standards, (b) determining whether a discrepancy exists between some aspect of the program and the standards governing that aspect of the program, and (c) using discrepancy information to identify the weaknesses of the program. (Provus, 1969, p.245)

Sociologists shared the approach: Greenberg (1968, p.260) wrote of evaluation as the 'procedure by which programs are studied to ascertain their effectiveness in the fulfilment of goals' and Brooks (1965, p.34) as '... the determination of the extent to which the program achieves its goals; the relative impact of key program variables; the role of program as contrasted with external variables'. Suchman's (1967, p.32) definition also contained this element, while Weiss declared that program objectives constitute the criteria against which results are to be assessed (Weiss, 1972a, 1972b).

Again, there is variation among those defining objectives as central to evaluation. Note that Brooks's definition, unlike that of Greenberg, suggested some diagnosis of what went on in the program – 'the key variables' – in its success or failure. His definition also hinted at elements in common with those who emphasize a scientific approach – in the contrast between 'program' and 'external variables'. Besides, each of these definitions implied that evaluation is an information-gathering process. Some have given special emphasis to this in their definition. Cronbach (1963, p.672) spoke of evaluation as 'the collection and use of information to make decisions about an educational program'; Hastings (1966), Forehand (1971), Frymier (1966), and Wiley (1970) have made similar statements.

Many associated evaluation with decision-making. Cronbach connected the collection of information in evaluation with its 'use' in making decisions (Cronbach, 1963, p.673). Forehand concurred, but

weakened the connection by saying 'this assessment presumably enters into decisions regarding the adoption, . . .' (Forehand, 1971, p.58). Similarly, Weiss saw evaluation as 'contributing to subsequent decision-making' (Weiss, 1972b, p.4). Elsewhere she asserted that the rationale of evaluation research is that it provides evidence on which to base decisions (Weiss, 1972a, p.328). On the other hand, neither Wiley nor Frymier included decision-making as an aspect of evaluation (Wiley, 1970; Frymier, 1966). It is noticeably absent from definitions put forward by Provus, Hammond, and Metfessel and Michael, and from definitions of many who defined evaluation as research.

In the 60s and 70s, there appeared in educational circles definitions of evaluation using the term 'systems'. As Glass (1968) has pointed out, since the term is currently used to refer to just about anything, it can hardly be said to have added a cutting edge to discourse about educational evaluation (see also MacDonald-Ross, 1973). Yet it is prevalent. It underlay Stufflebeam's (and so the Phi Delta Kappa, 1971, concept of evaluation as requiring evaluation of context, input, process, and product: the similar definition of evaluation presented by Alkin (1969a; b; 1972) and those at the UCLA Centre. Provus (1970) and Hammond (1969) presented their definitions of evaluation in terms which view education as a 'system'. It is closely allied to Tyler's emphasis on objectives, which extends the concept to include those of the institution (Forehand, 1971) as part of the educational system.

Cook (1970) summed up the view like this: a system has a basic purpose or function, which has to be accomplished. Yet since systems, like education, are complex, the system does not automatically correct variation and deviation that occur between the establishment of purpose and the accomplishment. Hence the system manager, in searching for the cause of that deviation, looks for data useful in defining the deviation and its seriousness, in establishing causes, in suggesting alternative solutions, and in choosing a solution. It is the function of evaluation to assist this process. Overlap with those who stress objectives as the standards for evaluation is obvious; less obvious is the overlap with those who see evaluation as an information-gathering process, and who stress its place in the decision-making process. Thus approaches like that of Cook put a good deal under a single umbrella, and the functions of evaluation become very broad.

None of the definitions above sees evaluation as placing a value on what is evaluated. Yet there are writers who see this as central. Scriven's reply to Cronbach's paper (Scriven, 1967; Cronbach, 1963) argued that evaluation *had to* involve the determination of worth or

value, that judgment was central to it, and that the judgment had to be the evaluator's. He defined evaluation as:

> ... gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings, and in the justification of (a) the data gathering instruments, (b) the weightings, and (c) the selection of goals (Scriven, 1967, p.40).

Among those who have concurred are Glass (1969), and Worthen and Sanders who jointly and separately have taken a similar position (Worthen & Sanders, 1973; Sanders, 1972; Worthen, 1968). Yet here, too, there are variations. Stake (1967) initially argued that evaluators were likely to be unwilling to judge a program, and suggested that instead judgment data be collected. More recently he has modified his position. In an article written with Denny, evaluation is said to include the task of gathering information about the nature and *worth* of educational programs in order to improve decisions about the management of those programs (Stake & Denny, 1969, p.373; with emphasis added, Stake, 1972).

Definitions so far noted draw on several traditions: science, sociology, administration and philosophy among others; but few reach into the humanities. Recently, it has been suggested that criticism has much to offer evaluation. Mann (1969) argued that evaluation could be equated with literary criticism viewed as a set of methods for analysis and disclosure. Eisner (1972, p.585) supported this, declaring that the function of criticism is to 'heighten one's perceptions of the qualities that constitute the work, ... the elements that constitute the whole and their interplay'; such criticism vivifies perception. Kelly (1973) developed the analogy between criticism and evaluation in a sustained manner. Something of the critic's approach is apparent in the Smith and Schumacher (1972) evaluation of the CEMREL Aesthetics project and Stake's recent statements about 'responsive evaluation'.

Evaluation defined as criticism rejects goals or objectives as the criterion, assigns a key role to qualitative judgment, and shows little concern with measurement, questions of control, and experimental design. But design of another kind has a central place. At least one writer has managed to combine, in his conceptualization of evaluation, the critic's emphasis on disclosure, the experimental emphasis on cause-and-effect relationships, and a fresh ingredient: 'responsibility'. To evaluate is to 'disclose the results of individual learning experiences, assign responsibilities, and look for causes of success and failure' (Wolf, 1971, p.38).

Different again is Rippey's concept of evaluation. All the concep-

tions of evaluation mentioned above suggest that evaluation is separate from what is evaluated. Rippey's concept of transactional evaluation, however, defined evaluation as a process for bringing about and sustaining changes within education. 'Transactional evaluation' includes protagonists and advocates of an innovation as well as a representative sample of persons likely to be adversely affected by the change. Its focus is on analysis of organizational dysfunction (conceived as threats to role stability imposed by the introduction of change on role incumbents) as well as improvement (Rippey, 1972, 1973). The evaluator's commitment is to making a program work.

Other definitions than those outlined above could be given: for example, the adversary model of evaluation or cost-effectiveness analysis. They are easily come by – but these are sufficient to show how the concept of evaluation has been greatly extended and blurred during the last fifteen years. Importantly, earlier conceptions in terms of measurement and objectives have not been simply superseded. A host of definitions and conceptions, so diverse that what is central to one may be omitted in another, have been put forward. It might be suggested that differences suggested by definitions do not extend beyond them. This is not the case. The many special terms that have been introduced to describe evaluation indicate the extent to which writers assert their view as different from those of others. In turn, these special terms create a sense of confusion. One cannot easily compare Scriven's ideas with those of, say, Stufflebeam or Stake. The relationship between Scriven's (1967) 'summative-formative' evaluation, 'intrinsic', 'payoff' and 'mediated' evaluation and Stake's (1967) use of terms such as 'antecedent', 'transaction', 'observed', 'contingency' and 'congruence' is cloaked by the language. Other terms have since been added to the pile, for example, Stufflebeam's (1969) use of 'context', 'input', 'process' and 'product' to serve 'planning', 'programming', 'implementing' and 'recycling' decisions. Others who have used still other terms include Hammond (1967, 1969), Provus (1969, 1972), Alkin (1969), and Rippey (1973).

Those introducing new terms have paid little attention to clarifying how they relate to terms already in use. The style has been to mention a connection in passing. For example, Rippey (1972) noted that his concept of transactional evaluation 'has similarities' to Scriven's concept of formative evaluation but he did not explicate. Stake (1967) commented on Scriven's and Cronbach's concepts of evaluation en route to introducing his own. Having selected from the former the idea that evaluation included judgment and, from the latter, that it included descriptive information, he then presented his own concept in a fresh set of terms. Yet the further task of making explicit the

connection between the position he had advanced and that of the other writers was left undone. By 1971 Proger declared that the plethora of terms was a sign that educational evaluation is playing a 'model building game' in which facilitating conduct of actual evaluations was no longer an object. More mildly, Weiss charged educational evaluators with creating a special language that blocked communication with others in evaluation in and outside education (Weiss, 1972a, p.13).

DIFFICULTIES WITH RECENT THEORY BUILDING

Both Weiss's and Proger's comments indicate that the discussions of educational evaluation, instead of clarifying the process and furnishing clear guidelines, have created fresh difficulties.

First, as indicated above, there is the number and diversity of ideas about educational evaluation. One writer, in an article significantly titled 'Confessions of a frustrated evaluator', declared:

> Anyone planning to deal with a theoretical or practical problem in educational evaluation these days cannot help but be overwhelmed by the massive and often conflicting body of literature ... with the possible exception of segregation, sex education, and student power, it is difficult to find an educational issue that has generated more rhetoric and greater controversy than the current concern for evaluation and the related issue of accountability. (Renzulli, 1972, p.298)

Second, the different ideas have been developed to varying degrees. Some writers content themselves with addressing a specific issue or with presenting a basic idea or principle. Cronbach's paper centred on the use of evaluation for course improvement (Cronbach, 1963); the concept of criterion-referenced tests was central to Popham's paper (Popham, 1969); Bloom has similarly dealt with mastery testing (Bloom, 1971): many have written about objectives (Atkin, 1968, Sullivan, 1969; Eisner, 1967a, b; 1969; Stenhouse, 1971a). Others developed what purports to be a complete 'model' or 'theory'. They present a detailed analysis and framework for evaluation, often complete with flow charts, summary tables and diagrams of how processes interlock. Those whose contributions have been of this kind include Suchman (1967), Scriven (1967), Provus (1971, 1972), Stake (1967), Alkin (1969), Stufflebeam (1969), and Rippey (1972). The evaluator is faced with a choice. He may be eclectic, in which case there is a considerable task in integrating ideas from various sources into some coherent plan, not to mention the prior difficulty of which ideas to choose (Sanders, 1972). Alternatively a single concept may be chosen. Something of the difficulties with this course of action can be gauged from the number of publications interpreting and summarizing the

more elaborate models, for example, Taylor & Maguire (1966), Glass (1968), Proger (1971), Farquhar & McCuaig (1972), Forehand (1971), Gephart (1972), Worthen & Sanders (1973).

Third, difficulties with the more elaborate pieces of theorizing in part stem from the way most have been developed. There is the lack of systematic examination and integration of ideas already noted (Tyler, 1969). Then too, those presenting ideas about evaluation do not clarify the basis or ground of the ideas, nor are the practical applications subsequently reported. For these reasons Lewy (1973) insisted that even the most elaborate conceptualizations or 'models' are more accurately referred to as 'speculative reflections' on evaluation, although their tone and language suggest universal applicability. It seems likely that authors have drawn on evaluations they knew and their experience in evaluation; but they have not specified the experience in sufficient detail and in such a manner that a second person would be able to work out the extent to which his evaluation overlaps or differs. Choice between concepts is not, at present, facilitated by information on evaluation studies used in developing the concept.

Fourth, if the evaluator attempted to solve the dilemma of conflicting theories by turning to accounts of evaluations, he would discover few available until very recently. Comparison of bibliographies to several recent articles on evaluation in the *Review of Educational Research* reveals that, out of a total of 273 references in four reviews written between 1969 and 1973, only 21 were devoted to accounts of evaluation studies (Baker, 1969; Westbury, 1970; Caro, 1971, and Sanders & Cunningham, 1973). Taylor and Cowley (1972, p.ix) justified the preponderance of theoretical articles in their book on the basis that this imbalance accurately reflected the state of the art: 'there are simply not many good pithy examples of evaluative activity in print'.

Fifth, features of evaluation treated as *de rigeur* in a particular conceptualization create difficulties for the evaluator in the field and may, in practice, be dysfunctional. Let us take the case of objectives. Most concepts of evaluation demand them. Yet evaluations have been done for programs that, even after the evaluation was complete, still had nothing like the kind of statement of objectives that Tyler had in mind, let alone the versions emphasized in the recent burst of evaluation models along the lines of the 'systems' concept (Alkin, Provus, Stufflebeam). (See for example MacDonald's (1971) evaluation of the Schools Council Humanities Project and Tawney's (1976) of Project Technology.) Another example comes from the conceptualizations that insist that the methodology of evaluation has to be 'scientific', i.e. has to be an application of the experimental designs advocated and laid

out so clearly by Campbell and Stanley (1963). In practice, evaluators have found themselves in situations where control groups are not readily available or become contaminated in the course of the evaluation (Sesame Street has had to deal with this problem) or where the central assumption that the 'treatment' has taken place must, itself, be examined carefully. Guba (1969) has been eloquent about the difficulties arising from conducting evaluations in settings far removed from the antiseptic environment of a laboratory.

Sixth, features not emphasized in conceptualizations sometimes emerge in practice as critically important. One of these is the role of the evaluator, interlocking as it does with the politics of evaluation (Cohen, 1970). Rossi suggested that the role of politics in evaluation has been seriously underestimated (Rossi & Williams, 1972). Grobman (1966) stressed that the relationship between project personnel and evaluator is critical; if it becomes soured, important information can be withheld. Another writer warned, '... the efforts of the evaluator are likely to be greeted with all the warmth of a husband who finds his wife in bed with his best friend' (Renzulli, 1972). On the other hand, if the evaluator becomes closely involved with the project, impartiality and credibility may be undermined. The point, for the moment, is that the role of the evaluator has been thrust into prominence from experiences in conducting evaluation studies. The formulation of it in terms of objectivity/subjectivity – roughly the terms in which Scriven dealt with it in 1967 – has been extended to include much more general issues of the strains that evaluation places on 'the researcher-practitioner relationship' (Rodman & Kolodny, 1965). Different aspects of that role have also been suggested, e.g. the relationship between evaluator and other groups for whom the evaluation is important (community, consumers, administrators and colleagues).

To sum up, each of these six points – the number and diversity of ideas; the varying degrees to which they have been formalized and elaborated; the lack of information about the evaluative experience on which concepts are based; the paucity of case-study accounts of evaluations; the features emphasized in the concepts which have been found in practice to be unimportant or to cause difficulties; and, correspondingly, the features important in practice which are not dealt with in the conceptualizations – create difficulties for anyone trying to understand evaluation or carry out an evaluation study. The resulting sense of confusion and dissatisfaction with educational evaluation is widespread. The constant flow of suggestions, conceptualizations, revision of techniques, and introduction of new ones over the past fifteen years implicitly suggests the extent of dissatisfaction. Explicit

expressions come from in and outside the field: administrators, curriculum workers, and evaluators. Many expressions are global attacks on evaluation; for example, an administrative viewpoint:

> Evaluation of the wrong kind, at the wrong time, and for the wrong reason has characterized too much of the current effort to appraise educational reforms. Meaningless evaluation is ruining the cutting edge of educational innovation. (United States. National Advisory Council on Education Professions Development, 1969)

Curriculum developers have spoken of evaluation in similar terms. Wilhelms charged:

> It (the present evaluation activity) is equally to be condemned for the narrowness of its focus, because in its gross exaggeration of the more mechanical easier-to-measure features of education, it virtually blots the broader more fundamental objectives out of sight. The end result is not simply bad evaluation; it is distorted teaching and learning. (Wilhelms, 1967, p.15)

An evaluator with fifteen years experience asserted:

> I have come to the conclusion after much soul searching, that evaluation is at present the major disaster area in education ... most evaluation is a fraud, or at least a dangerously misplaced effort ... the results of our evaluation are superficial and tenuous ... (Macdonald, 1974, p.3)

Each of these three charges made about evaluation is serious, yet the form of the charges makes the criticism less useful than it could otherwise have been. The National Advisory Council's comments asserted failure of evaluation in global terms. The complexity of approaches used is not admitted, nor are the elements of an approach contributing to failure distinguished. Wilhelms equated what sounds like a particular approach to evaluation with all evaluation. Macdonald's comments may have expressed the feeling of many, but they did little to clarify the aspects of evaluation or concepts which contributed to it.

Some writers have attempted to articulate criticisms more fully. Reviews of evaluation published since 1966 in the *Review of Educational Research* have repeatedly pointed to the need for a variety of 'model' evaluations and for a framework that brought together the different views. In 1966, Abramson concluded that evaluation had:

> failed to provide even a minimally acceptable way of responding to practitioners' needs. (There is) a continuing paucity of studies which can serve as models for curriculum research. (Abramson, 1966)

Others repeated the criticism and made additional ones. Westbury's (1970) review went further. He concluded that a major problem in curriculum evaluation was this:

> Two separate, though interrelated analytical problems must be faced: curriculum must be conceptualized in such a way that it no longer carries

> the connotation that it is a unitary notion, often a treatment; evaluation must be seen in ways that permit the development of *sets* of methods and criteria, so reasoned judgements, appropriate to *all* senses of curriculum, become possible. [sic] (Westbury, 1970, p.245)

Westbury is arguing for frameworks in curriculum and evaluation that do not treat the phenomena as unitary; that, while providing a focus, preserve pluralities.

Other evaluators have been, in some cases, more successful in articulating specific criticisms. One of the most influential of these, Guba, shared the conclusion of Wilhelms, Macdonald and others that evaluation had 'failed' and detailed specific issues in that failure. Guba (1969) cited the following 'crucial lacks' in educational evaluation: adequate definition, adequate theory, adequate integration of evaluation with decision-making, adequate dealing with values and criteria, clear differentiation of administrative levels (micro from macro), mechanisms for reporting evaluative information, and trained personnel. The first two 'lacks' especially relate to concepts of evaluation. The next four list particular weaknesses apparent in practice which Guba claimed are omitted or inadequately addressed in the theorizing. The last point is of a different order, not susceptible to theorizing. Since most of the points made in that article are repeated in the first chapter of the Phi Delta Kappa (1971) volume, it seems reasonable to conclude that there was solid support for Guba's views among educational evaluators. Guba's criticism suggested that concepts of evaluation were ill aligned with practice. Other evaluators supported his conclusion, and suggested that the demands of conceptualization have themselves become a dominating preoccupation:

> It is possible that they (evaluation theorists) have in fact overanalyzed the process . . . Have, in fact, these analyses departed from operational utility, at least in the sense that the practitioner would not know what to do with them? (Finn, 1969, p.18)

or, again:

> In a conceptual hierarchy, every concept eventually rises from a specific level of applicability to a general level of functional incompetence through logical analysis. (Proger, 1971, p.293)

Educational evaluation, like curriculum, has shown signs of being inveterately theoretic (Schwab, 1970).

EFFORTS TO CLARIFY THE CONFUSION

The evaluation field is a lively one. While it has not yet achieved definite focus, since 1970 there have been attempts to clarify the confusion and to pay more attention to practice. Collections of writing

about evaluation are appearing in increasing numbers (Caro, 1971; Rossi, 1972; Weiss, 1972a; Taylor & Cowley, 1972; House, 1973; Rippey, 1973; Worthen & Sanders, 1973; Payne, 1974; Popham, 1974; Walberg, 1974). In addition, there have appeared two reports on conferences on evaluation (Wittrock & Wiley, 1970; Apple et al., 1974). The collections do provide perspective on evaluation; they generally place side by side various concepts of evaluation and articles discussing related issues. However, usually there is little analysis of the views, few explicit comparisons, and little attention to differences in assumptions and conclusions. The books make readily available key readings in the field and they provide, in some cases, a sense of what someone perceives as key issues in evaluation. At their best, they present the would-be evaluator with a 'tree full of owls' (Weiss, 1972a) whose counsel, hopefully, is wise and sufficient.

Worthen and Sanders (1973) do go a step further. Their book not only collects within one pair of covers major conceptualizations by educational evaluators, but does so in a way that attempts to point out what is distinctive in each.

The third chapter brings together previously scattered writings by Cronbach, Scriven, Provus, Hammond, and Stufflebeam, among others. Eight of the views presented are then summarized in a matrix which highlights similarities and differences on selected aspects (definition, terminology, and so on). Yet no attempt has been made to consolidate where views have essentially the same elements, nor to consider implied differences and similarities (Gephart, 1972). The volume also gestures at the relationship between concept and practice. Each author's concept of evaluation is 'applied' to the same evaluation project: a classroom teacher with a new curriculum. The applications are stilted and superficial (Carlson, 1974).

Other collections have presented case-study accounts of evaluations (Walberg, 1974; Rippey, 1973; House, 1973; Caro, 1971a; Payne, 1974; Taylor & Cowley, 1972). One consists of little else (Schools Council, 1973). In addition, a few case studies have been published in journals (for example, an issue of *Viewpoints*, July 1972; Welch, 1972; Lindvall, 1966; Grobman, 1968). The number of accounts is steadily increasing, which suggests that evaluators are observing and studying evaluation-in-action, and addressing themselves to the relationship between concept and practice. The case studies vary: some are close to evaluation reports (particularly those in Taylor and Cowley, and Payne). While these do convey something of an evaluation in action, usually it has to be teased out of them, since evaluation reports

> usually tell little more than that the work proposed was completed, that the complaints of the staff were justified, and that there were greater

> differences within groups of students (or schools, or curriculums) than there were between groups. Most evaluation reports give only the participants some notion of what occurred; the outsider gains little insight. (Stake & Denny, 1969, p.370)

Most represent reflections at the end or some mid-point of a project for which information has not been carefully collected and analysed. They do, however, give a sense of evaluation-in-action that was previously missing. When the case studies are juxtaposed with theoretically oriented discussions of evaluation, differences become apparent.

Naturally, the writers of the case studies are themselves influenced by concepts of evaluation, some adhering quite closely to a particular view. Those published in Walberg's book are there partly to exemplify his view; the same is true of those in Rippey's, or the interlacing references to evaluation work in Provus (1972). However, similar material in House's volume has been chosen to illustrate not the views of a particular theorist but contrasting accounts of evaluation politics. Other material is only loosely related to a particular position or concept of evaluation, e.g. the Schools Council case studies. Besides, evaluators have, presumably, familiarity with more than one concept of their task and this, together with the constraints of the particular situation they had to work under, is depicted in their accounts.

More difficult to estimate than the influence of a particular concept is the influence of what is publishable in journals on the case studies that have become available. Until recently there have been few vehicles for publishing this kind of material. The Harvard Project Physics manuscript on the evaluation has not found a publisher; a shortened journal article on the subject recently appeared in the *American Educational Research Journal* (Welch, 1972). Several published studies have strong 'research' components – see for example, the four in the *Journal of Educational Measurement* (1966), that of Welch, above, and most of those listed in Walker and Schaffarzick's (1974) comparison of curricula.

3

Development of the Characteristics

An evaluation is an event in the life of an institution. It can, of course, be a continuing monitoring operation which extends over a long period, even to the total life of the organization. More typically, it is something phased in and out of the continuing history of an organization. An evaluation moves from beginning to end, and is full of doing, of action.

Evaluation studies differ in hundreds of thousands of particulars. They take place in a particular context and are concerned with a particular project, however diffuse or complex. An evaluation is carried out by particular persons or groups at a certain time and place. An evaluation is time bound and budget bound, constrained, too, by a thousand other contextual factors. Interacting with these are the skills and idiosyncrasies of evaluative and project staffs. No doubt in evaluations there are personality clashes, no doubt information is deflected, decisions made differently because it's a rainy day or there was an accident on the way to work. There is no reason to think that the particulars of an evaluation and its idiosyncrasies do not influence and shape the project. But it is reasonable to think that these particulars can be subsumed under some abstraction so that in another evaluation study, with quite different specifics, some parallel is seen. Such abstractions are referred to as characteristics of educational evaluation.

PURPOSE OF THE CHARACTERISTICS

The characteristics of educational evaluation were developed to provide a framework for facilitating comparisons within the field – particularly comparisons between evaluations. Comparisons can also be made between evaluations and concepts, and between one concept of evaluation and another. The characteristics are to serve as the medium, indicating the comparisons to be made. Someone making a comparison would use the characteristics to point out aspects of the evaluations particularly deserving of attention, that is the non-trivial

features that distinguish one evaluation from another. Someone making a comparison would have a record, not of the characteristics themselves, but of concrete details, including the idiosyncratic and individual. The characteristics, then, are to function as pointers or questions about an evaluation, answers to which require delineation and analysis of many concrete instances. Thus the characteristics are abstractions.

FORM OF THE CHARACTERISTICS

Abstractions sum up a pattern or feature, as distinct from what is individual, idiosyncratic, or particular in a phenomenon (Robinson, 1950, p.197). Thus, that one becomes aware of a particular tree out the window is not usually called an abstraction. But when we become aware of the similarities and differences in form between, say, a flowering plum and a sycamore, *and* the discriminanda are encoded symbolically (including verbally) so that it can be applied to the next thousand trees we see, then we have an abstraction. An abstraction may be analysed. Having named a play a tragedy, so forming an abstraction of some general characteristic of it, we may ask what a tragedy is. The response to that question involves some analysis of tragedy during which the complexity of elements that contribute to the characteristic becomes clear. To return to an earlier example, one can be aware not just of a particular tree but of it as a tree, and the many features that distinguish it from the grass in the field in which it stands. Awareness of that can be quite distinct from the analysis of the tree which makes apparent the complexity of its relationships, for example, that some trees have large leaves and sway in the breeze while others are fine-needled and stand stiffly.

To sum up, it is possible to identify two aspects of a characteristic. The first is the general feature or characteristic to which a name is given; the second, the analysis of the characteristic. In the context of educational evaluation, this would suggest first, that particulars of evaluations have to be subsumed under a set of characteristics and, second, each of these characteristics has to be analysed to identify the elements and complexity of relationships which make up that characteristic.

There are, however, several difficulties in forming abstractions. First, there is no mechanical way of accomplishing the task. 'Like artistic creation and scientific theorizing, it is a groping affair' (Robinson, 1950, p.175). An abstraction has to be conceived, an idea formed as to its analysis, and then the two have to be tried out in some way. There is no sure method for conceiving abstractions, developing a sense

of the complexity subsumed by one, or of trying out the ideas once conceived. It appears that abstractions have to be groped towards, considered in relation to information of different kinds about evaluative activity, and then a judgment made. A second difficulty follows from the first. Since the procedure is a groping affair, there is no clear-cut point of completion.

CRITERIA USED IN DEVELOPING THE CHARACTERISTICS

A very large number of characteristics could – granted sufficient imagination – be developed. The possible number of abstractions exceeds the number of evaluation studies. How then does one select? Having perceived a characteristic, why include it in the cluster and not another? Criteria commonly applied to theorizing apply to this project. First, the abstractions should preserve pluralities. Second, overlap between categories should be minimized. Third, characteristics should be consistent with each other in terms of congruence and level of generality. Fourth, characteristics should be relatively easy to apply. Fifth, the final listing should be reasonably comprehensive. Each of these criteria gives grounds for selecting among the possible abstractions, limiting the number. The use made of each of these criteria in developing the set of characteristics is outlined below.

To facilitate comparisons across a wide range of evaluative practice and theory, the abstractions must preserve pluralities. Abstractions that defined a single unitary activity without variation or a single correct form of evaluation are not included in the list of characteristics. For program 'objectives' to be included as a characteristic, the term could not be equated with a single usage (such as the expressive objectives suggested by Eisner, 1969). Instead a range of alternative practices and concepts had to be included. The abstractions must reveal the options. This gives critic and evaluator alike a perspective on what might have been done in an evaluation and in what other terms the evaluative study could have developed, a sense of the choices made and a view of the style of a single evaluation study or conceptualization of evaluation (Schwab, 1970; 1971).

Redundancy is undesirable, just as economy is desired. Although some interconnection between characteristics may be desirable, overlap has been minimized. Redundancy is minimized when the abstractions or characteristics are conceptually and functionally distinct. Thus a program's 'goals' and its 'diversity of form' may form two distinct characteristics. Yet some redundancy has been retained. The extents to which an evaluator 'interprets' and 'judges' are conceptually dis-

tinct, yet functionally they overlap; laying out data is likely to form part of both.

There is a band of generality within which the characteristics should fall. Characteristics should be sufficiently abstract to subsume particulars of evaluations, and yet sufficiently clear-cut to enable their use in examining examples of evaluations or conceptualizations. In these terms, a characteristic like 'data analysis' is too specific. One could easily determine the application of 'data analysis' to an evaluation, but there are studies using approaches that precluded data analysis. For example, a recent volume of evaluation case studies includes a narrative account of a program in a literary style (Brauner, 1974). Yet a characteristic like 'complexity' is too general. Evaluations do vary in complexity. It is difficult to think of an aspect of an evaluation which would not be subsumed under 'complexity'. Some possible characteristics are not included because they are peculiarly difficult to apply. Sometimes a characteristic may be so abstract as to be impossible to apply. At other times, difficulties of application are not related to the level of abstraction. For example, 'purpose of the evaluator' in conducting his study is an interesting possibility for a characteristic. Evaluations are presumably conducted for some purpose. Yet in trying to identify what is included in the characteristic, there is the difficulty of ascertaining an evaluator's intentions. To do so with any certainty requires knowledge about what goes on in the evaluator's mind. In contrast, a characteristic like 'utilization' refers to information that is likely to be accessible and so permits appraisal of the extent to which it applies in a particular evaluation.

Finally, the characteristics as a set should be reasonably comprehensive. If an evaluation study is described using the whole set of characteristics, the resulting profile of the study is fairly comprehensive. One may readily think of details that have been omitted, but there seems to be no major aspect of the evaluation not included somewhere. Two procedures have been used to help assure that the set does include most of the key characteristics of educational evaluation. The first is a wide survey of the ideas of those writing about evaluation in and outside education. The second is the testing of the set of characteristics on actual instances of evaluation. A sample of the resulting profiles has been set out in Chapter 5. There seem to be no major aspects included in the case-study materials and not represented in the descriptions developed by using the set of characteristics. The set is not exhaustive in a strict sense. It seems impossible to attain that limit, given the present state of development of evaluation and, more particularly, the extent to which the field is still going through a period of development. New features are likely to become part of the area

in future and others to cease to be of importance. For example, the evaluator's analysis of possible causal processes included in the program may become a major feature of evaluations in the future. Certainly some are now arguing for its importance. (See Weiss, 1972b, pp.49–53; Scriven, 1974, pp.68–84.) The list of characteristics includes analyses of processes without specific references to causality – the latter could be added later. But use of the above criteria on their own still leaves more possible characteristics than could be of utility in making comparisons. The set is thus limited to 'key' characteristics, that is those that distinguish evaluation from other education activities and those that point to variations in style.

In describing an evaluation numerous details are used: budget, size of staff, time line, sources of information, size of sample, and so on. These categorizations themselves constitute abstractions in the terms described earlier, yet they seem relatively insignificant features of an evaluation study. These features are not peculiar to evaluation; some are equally applicable to research projects of the most 'pure' or 'basic' variety (for example, Patterns of Declarative Utterances of Seventeenth Century Metaphysical Poetry), and in some cases (budget and sample size, for example) to a whole array of business organizations. Such categories, then, do not appear to be useful 'distinguishing marks' (Cassirer's (1970) phrase) for educational evaluation per se, although they could well be useful in discussing organizations.

Some possible characteristics tell us relatively little about the evaluation. Budget and sample size do vary from one evaluation to another, sometimes dramatically as evidenced by comparison of the Coleman (1969) study with a School District's Title I study. Yet the two features formulated this way do not subsume a complex of interrelations. They have less to offer than abstractions like 'role of the evaluator'. Budget, staff, instruments, sources of information, reports are abstracts of the components of an evaluation (analogous to the wheel, pedal, frame of a bicycle or the stage, set, props, cast of a play). As such they tell us relatively little of differences in style between one evaluation and another, although they do, of course, contribute to those differences. Similarly, divisions of evaluation studies into stages, while abstractions, present us with another set of components of evaluation as a process unfolding through time, analogous to the acts and scenes of a play. Characteristics that function as heuristics, or have the potential for doing so, have been included over those that do not.

RELATIONSHIP OF THE CHARACTERISTICS TO THEORY

The framework of characteristics does not aspire to be, nor does it

constitute a theory, in the sense of a highly structured body of generalizations that systematically employ concepts offering some approximation of axiomatic treatment (Braybrooke & Lindblom, 1963). To redigest all the current thought into a single overriding theory would not preserve pluralities or allow for the complexity of evaluation studies, although it would be in a tradition that dates back to Plato and Aristotle. Schwab articulated his objection to this approach:

> Quasi-platonic accounts assert the merely contingent character of any and all joints, seams, or separations in nature and emphasize the potential diversity of subject matters, but denigrate pluralities of theories based on such diversities and obscure their usefulness by treating all contingent joints and separations as merely imperfect starting points of intermediate inquiries to be used only as stepping-stones to an ideal, wholly unified theory of everything, of all the world, seen as an englobed unity. (Schwab, 1971, p.510)

A more radical objection to a single overriding theory is that definition in terms of inherent principles assumes that most of our abstract terms are univocal, when our daily use of language demonstrates ambiguities and the extent to which a single term has a range of applications (Robinson, 1950, pp.152–156).

Yet the aspiration to scientific theory is shared by some of those conceptualizing educational evaluation. The concluding section of the Phi Delta Kappa (1971) volume on evaluation reviews the extent to which preceding sections represent an adequate theory of evaluation. Noting that the decision is a task for the 'philosopher of science', the committee argues that a theory must be a coherent set of hypothetical, conceptual, and pragmatic principles. The terminology employed should be coherent. The theory should promote deduction of testable hypotheses which reflect direction on the validity of the theory; it should be comprehensive and internally consistent; it should permit the derivation, operationalization, and execution of evaluation designs which are scientifically adequate; and it should be a general guide for a wide range of evaluation problems (Phi Delta Kappa, 1971, pp.328–330).

The extent to which the current stage of development of evaluation is suited to the kind of theorizing just described is debatable. Kuhn has argued:

> In the early stages of development of any science, different men confronting the same particular phenomena, describe and interpret them in different ways. (Kuhn, 1970, p.18)

Accounts of evaluation studies enable us to see some of the gaps in the concepts that are being advanced and some of the areas in which

they conflict or overlap. They make tangible for us the partiality and incompleteness to which theorizing in the social sciences is particularly prone. Each of the concepts, particularly those that are more comprehensive in appearance – like those of Stake, Stufflebeam, Provus, offers a different perspective on evaluation, highlighting some aspects and subordinating or excluding others. Each presents a certain perception of what is important in evaluation, and usually in a manner which suggests that there is only a single correct answer. Yet in education it is important to bear in mind that concepts are also vehicles for persuasion (Scheffler, 1960; Apple, 1974; Robinson, 1950). In effect, such concepts as those of Stake, Weiss, Cronbach, Stufflebeam, and Scriven attempt to persuade others to look at evaluation in a particular way by stressing certain aspects: what we tend to forget is to look for what is being excluded. Often such concepts are prescriptive rather than descriptive.

The framework presented here is not intended to constitute a theory, nor does it function as criteria by which to judge an evaluation as 'good' or 'bad'. Scriven has developed such a set of criteria for judging the merits of an evaluation, including whether the field trials of the program were 'perfectly typical', whether performance data were collected for 'all relevant consumers', and whether the statistics represented 'flawless analysis' of the data (Scriven, 1974, pp.7–33). Instead, the characteristics of educational evaluation are intended to identify the key aspects of evaluation to be used in developing a description of an evaluation. They bring into focus important aspects of the phenomenon, but they do not on their own indicate whether the study is 'good' or 'bad'. Use of the characteristics may indicate that quantitative analysis plays an important part in an evaluation, but not whether those analyses were well done. The function of the framework is not to define what 'ought to be', but rather to indicate the range and complexity of the evaluation.

4

Presentation of the Characteristics

The key characteristics of educational evaluation set out in this chapter are arranged in four groups. The first group of characteristics relates to the program under evaluation; the second, to the stance the evaluator adopts in conducting the study; the third, to the questions to which the evaluator seeks answers; and the fourth, to the evaluation design, information collection, and reporting.

Other groupings of the characteristics are possible: for example, questions to be investigated, procedures used in the study, and interpretation and communication. The one chosen provides a convenient framework which includes all of the characteristics and which corresponds to major aspects of evaluation studies.

Program Aspects

Characteristic I:	Program's Stage of Development when the Evaluator is Appointed
Characteristic II:	Program's Openness to Revision
Characteristic III:	Program Uniformity from Site to Site
Characteristic IV:	Specificity of Program Objectives

Evaluator's Stance

Characteristic V:	Evaluator's Independence
Characteristic VI:	Evaluator's Orientation to Value Judgments on Program Merits
Characteristic VII:	Evaluator's Orientation to Utilization of his Work

Questions Asked in the Evaluation

Characteristic VIII:	Attention to Program Goals
Characteristic IX:	Generalizability

Evaluation Design, Information Collection, and Interpretation

Characteristic X:	Formality of Control
Characteristic XI:	Emphasis on Comparison
Characteristic XII:	Scope of Information Collected
Characteristic XIII:	Emphasis on Interpretation

Figure 1. Summary of the Characteristics

Each characteristic has two parts: a brief title that sums up the central concept, and a list of elements or factors that contribute to that central concept. (A summary list of the titles is given in Figure 1.) While all characteristics are relevant to any evaluation study, individual elements within a given characteristic may or may not be applicable to a particular study.

The first group of characteristics has to do with the program evaluated. Although the evaluator may be able to influence these, to a large extent they function as constraints with which he has to work. Programs vary in complexity, in the size and kind of population they serve, in duration, in the way they are organized, and the extent to which they are innovative. For example, a program may be tightly structured and unified (e.g. Montessori materials). It may also be loosely structured, assuming diverse forms in the hands of different users. In either case, there are consequences for the evaluator:

> It is one thing to evaluate the effects of a small, short term, specific, well-defined program, such as a training film. It is a far different and more difficult matter to evaluate the effects of a national antipoverty program, with its diversity of methods, actions, and goals. (Weiss, 1972b, p.4)

CHARACTERISTICS RELATED TO THE PROGRAM

Characteristic I : Program's Stage of Development when the Evaluator Is Appointed

The characteristic sketched here inquires about the timing of the beginning of the evaluator's work in relation to the development of the program. The importance of the characteristic lies in (a) the possible effect the stage of program development has on the evaluation study, and (b) the possible contribution the evaluator may make to the program.

If an evaluator starts work well into the life of a program, several aspects of his or her work are constrained. For example, some data are impossible to collect. Complete baseline data prior to the commencement of program operations will be unavailable, 'adequate information about operations can almost never be obtained retrospectively' (Riecken, 1972, p.93). The evaluator who arrives on the scene early can collect data on the state of affairs prior to the action phase of the program. He may be able to introduce systematic variations in treatment, involve program personnel in the design of the evaluation, and develop in them an 'experimental attitude' as distinct from blind commitment to the program (Riecken, 1972, pp.93–101). The

evaluator may also be able to contribute to the program by clarifying objectives, encouraging program personnel to use methods of proven effectiveness, and pointing out likely difficulties in operations.

An evaluator may also arrive too early in a program's life. He may be there to 'help dig the foundations' (Herzog, 1959, p.84) but have nothing to evaluate. Premature evaluation may curtail the trial and error of program development too soon, and lead to definition and structure before program personnel are clear about what they are attempting to achieve:

> An evaluation should not be undertaken until an activity or program has had enough time to prove its possible effectiveness. Obviously, the program has to have become operational before it can be evaluated. (Suchman, 1968, p.55)

Yet the writer goes on to warn:

> The evaluation should occur while there is still time to apply the results to the operation of the program. (Suchman, 1968, p.56)

Much has been written about the stages of development in a program. Most lists are variations on the stages suggested by Alkin, Suchman, Guba and Clark, and Gideonse. Alkin (1969a, pp.1-6; 1969b, pp.2-7) suggested five stages: systems assessment, planning, program implementation, program improvement, and certification of the program as a completed product. Suchman (1967, pp.134-142; 1970, pp.56-64) suggested four stages: research, planning, demonstration, and operation. Guba and Clark (1965; Guba, 1968) identified: research (with needs assessment added to Suchman's meaning), development (including invention and design), diffusion (including dissemination and demonstration), and adoption (including trial of the program, its installation, and its institutionalization). Gideonse (1968), discussing Guba and Clark's proposal, settled on three stages: research which furnished the knowledge base for planning, development, and production (in which the program was demonstrated and then made fully operational).

First, comparison of the suggestions of the four authors makes it apparent that each identifies a stage of program development prior to implementation, which may be broadly termed 'planning'. Within the planning stage, each distinguishes some assessment of the program's basis in terms of 'needs' it will attempt to meet (Alkin), resources and attitudes of likely clients (Suchman), the development of knowledge required for subsequent planning (Gideonse). Guba speaks in similar terms and, with Gideonse and Suchman, admits basic research to a place. Each author also refers to program invention and design. No author makes a sharp distinction between planning and

development at this point. Even Guba's substages of invention and design are presented as being carried out hand in hand. Thus it is reasonable to identify a planning stage in the development of a program that may include: (i) basic research, (ii) assessment of resources, need for the program, etc., and (iii) inventing and designing the program. In (iii) program personnel would develop goals, procedures, and policies for the program; select resources, equipment, facilities to be used in the program, and, in curriculum projects, possibly design materials and strategies for teachers as well as students. Programs may differ on the sort of work done during the planning phase, the amount of time spent, and the extent to which some planning is carried out concurrently with implementation. Yet the various phases seem distinct.

Second, each author refers to a stage in which the program is first implemented. Alkin distinguishes two stages: implementation and improvement. Suchman and Gideonse identify a demonstration phase, within which Suchman distinguishes three subsections. Guba lists 'program trials' as the first stage in diffusion. Common to each of these concepts is that of a program being tried out and altered. Suchman subdivides the stage of program implementation into three: pilot, model, and prototype. In the pilot stage, the program functions on a trial-and-error basis with much flexibility in which to introduce alternative procedures and changes. In the model phase, the defined program is tried out under controlled or even under laboratory conditions, while in the third, a prototype program is tried out in a natural setting. When a program is implemented as a prototype, the emphasis is on application in a practical setting of a reasonably stable program (Suchman, 1968, p.62).

Third, each of the authors identifies one or two further stages in program development. Alkin speaks, once more from the evaluator's stance, of program certification; that is, of the program at a finished stage in which no more amendments are expected. He does not carry development of the program through to a stage when it becomes part of an institution. However Gideonse, Suchman, and Guba and Clark all do.

When a program is adopted by a new group of users or located at a new site, there ensues a period in which the program becomes established, the new staff become used to administering it, and adjustments are probably made. Local staff may be trained; changes may be made to equipment and materials used; the program may take some time to settle into a routine. An institutionalized program is one that is well established and in which development, while it may still be continuing, is minor. The program has been absorbed into

the running of a large institution and is no longer regarded as new or different but is simply part of the organization. Compared with a program being adopted by new users, an institutionalized program has greater stability. The familiarity of those administering an institutionalized program may even reach the point where they become largely unconscious of specific procedures, steps in treatment, strategies, and goals.

To sum up, four major stages in program development have been distinguished: the first dealing with planning, the second with implementation, the third with program adoption, and the fourth with institutionalization. In each of the first two stages further phases have been noted. Program planning may include basic research, needs assessment, and invention and design of a program. Program implementation may include three phases: trial and error, the program as a model, and the program as a prototype. Each stage is set out in summary form below, starting with the earliest and moving to those that come later in the life cycle of a program. It must be emphasized that while A–D represent stages that may be distinguished from each other and through which some programs may pass, it is likely that some programs will bypass certain stages, stages will overlap in some programs, and stages will vary in the proportion of time they take from one program to another. The process is not always lineal and fixed in duration. At each stage, the decisions that have been made already by program personnel, and hence the conditions under which the evaluator works and decisions in which he may participate, differ. The characteristic is summarized below:

A Planning:
  1 basic research
  2 needs assessment
  3 invention and design

B Implementation:
  1 trial and error
  2 model program
  3 prototype

C Adoption

D Institutionalization

Characteristic II: Program's Openness to Revision

In the distinction between summative and formative evaluation,

Scriven (1967) suggested that program revisions are made primarily during the formative or developmental stages of a project for purposes of improving it. The summative/formative distinction has also been used to suggest that there is a point where a program becomes fixed or set, a product on which final judgment is to be passed without considering ways of improving it further. It has been assumed that at this point the program is no longer open to revision.

> Many programs, however, are never 'finished' in the sense that a curriculum is finished, and continual modification and adaption will be necessary both at the original site and in other locations that use the program. (Weiss, 1972b, p.17)

Revisions may extend well beyond the development phase. Even published curricula have been revised. Thus, the distinction between finished products (that is, not open to revision) and those still in the making is not clear-cut. Ball, commenting on summative evaluation, makes the point thus:

> The study should also be sufficiently sophisticated to allow the evaluator to point to the program features that seem to influence success and failure. That is, good summative evaluation may well have formative implications. (Ball, 1975, p.408)

Yet, a program's stage of development partially shapes the extent of possible revisions. While a program exists only as a plan or a blueprint, extensive revisions are possible. When a program has only recently moved into action, there may still be considerable room for revision, although sometimes the form of the program may be so fluid as to make revision difficult. Once a program has become established, those conducting it may be interested in modifications, but may have invested so much that comprehensive rewriting is difficult. Yet once an established program is disseminated to a new site, minor revisions in order to get it running smoothly are likely.

Revisions are more likely when those carrying them out are both available and interested in doing so (Rodman & Kolodny, 1965). Program directors or those currently using a program are readily available and in a position to make revisions. Where those with the power to alter a program are not directly associated with it – as in the case of policy-makers – changes are unlikely. But those with the power to make revisions may not be prepared to do so. Program directors and adopters may view the program as fixed. Policy makers may be concerned only with its success and not in tangling with possible modifications (Weiss, 1972b).

There are difficulties in ascertaining whether program developers or adopters are prepared to revise a program. Evaluators are often

hired or imposed on programs as a condition of funding, thus program directors' interest in revisions based on the evaluators' work may well be lip-service (Longood & Simmel, 1962). Further, a director may be prepared to make revisions and not be able to do so. Possible constraints are legion: political climate, time to carry out revisions, costs, skill, and so on. Some indication of whether those in a position to revise a program are interested in doing so may be gathered from statements about the program and statements to the evaluator.

In general terms, a program includes three aspects: goals that set out program aims or objectives; students, staff, materials, media and equipment that constitute what is used in the program; and activities, treatments, and procedures that make up the program (Stake, 1967; 1969). Particular programs may lack one or the other of these aspects (for example, stated goals); or place much less emphasis on one (for example, equipment) than other programs. Some or all of these aspects may be open to revision. Program directors may regard goals as fixed, but be prepared to revise other aspects. Similarly, they may be committed to using certain materials but be prepared to make revisions in how they are used. In some situations, all three aspects – goals, what is used in the program, and what is done – may be open to revision. Although program personnel may be prepared to revise several aspects of a program, alterations may be limited to specific details. For example, program personnel, finding that goals are being misunderstood, may rewrite them in other terms but not be prepared to abandon them. Thus specific details may be modified but the general framework of the program is not open to revision. Revisions may also extend beyond specific details to affect the organization and policies of the program. On occasion, a program's rationale and structure may be altered. To do so would constitute a drastic revision.

To sum up, there are several elements in a program's openness to revision: who is to make revisions and their availability, the program's stage of development (still being planned, just being implemented, being installed at a new site, well established), the extent of possible revisions (every aspect of the total program, several selected aspects, a single aspect, none at all), and at what levels (specific details or particulars, organization and policies, general framework and rationale). Each of these is set out in summary form below:

A Program's stage of development:
- 1 planning
- 2 initial implementation
- 3 adoption
- 4 well established

B Who would carry out revisions:
  1 program developers
    (a) available
    (b) interested
  2 users or adopters
    (a) available
    (b) interested

C Scope of revisions:
  1 every aspect of the program
    (a) goals
    (b) what is used, e.g. staff, equipment, materials
    (c) what is done, e.g. activities, treatment, procedures, methods
  2 selected aspects
    (a) goals
    (b) what is used
    (c) what is done
  3 single aspect
    (a) goals
    (b) what is used
    (c) what is done
  4 no aspect is open for revision

D Level of revisions:
  1 specific or particular details
  2 organization and policies
  3 general framework, structure, and rationale

These may be combined to indicate the extent of a particular program's openness to revision although each element may not be relevant to a particular program. For example:
A program is not open to revision where directors of a well-established program take the position that either it is to be accepted as it is or not at all.
A program is open to some revision where developers, having only recently implemented it, are interested in some modifications in activities and methods that are not working satisfactorily.
Extensive revisions to a program are possible where it is still being developed and those able to carry out revisions are available and interested in revising several aspects of it.

Characteristic III: Program Uniformity from Site to Site

A uniform program is one which remains the same irrespective of con-

text. A uniform program would be exactly reproducible wherever it was placed. Such a program would be standardized in every respect. Such uniformity depends on control, achieved in the first place through careful specification and design, then maintained through systematic monitoring. Specification establishes uniformity; monitoring serves to maintain it by preventing variations in use. Variations may be minimized by, on the one hand, supplying resources to be used in the program and, on the other, training staff who will administer the program. The Intermediate Science Curriculum Study, for example, is not only prepared to supply chemicals and apparatus (including a box of rubber bands) but even a trolley to hold the equipment. Similarly, when program personnel conduct staff training sessions, then sources of possible diversity among instances of the program are less likely than when staff work solely from specifications or guidelines. Finally, where a program gives no specifications and no supplies, and does not train staff, versions of the program will differ.

The overall program structure and organization also contribute to uniformity and diversity. In general, uniformity is weakened where choices are permitted. As more choice is introduced into the situation, uniformity and standardization are lost and variation and diversity take over. When a program is so organized and structured as to permit those using it to select among specified alternatives, there will be noticeable variation in the forms of the program. To take the example of the program with four subsections, if the users may work with any two, there will be variation. Yet the range of variation is limited by the number of alternatives available. When a program, either in entirety or for certain sections, permits users to develop their own version, enormous diversity can develop.

Each of the forms of control so far mentioned contributes to *establishing* program uniformity. If program uniformity is to be *maintained*, the program-in-action needs to be monitored by some central agency. Even if a program is carefully specified, once it is in operation variations may gradually develop – or be deliberately developed to meet local conditions. Unless the program is monitored, variations may pass unnoticed. Where there is provision for control of program variations, it seems reasonable to expect uniformity. Where these are absent, variations and diversity may be expected.

Program uniformity is, however, just that, the standardization of a *program*. Uniformity is affected by the program itself. Where a program consists wholly or for the most part in tangible concrete objects, there are reasonable grounds for expecting it to be uniform. For example, if a program relies heavily on slides, tapes, films, videos, booklets, and so on, uniformity from one context to another is reason-

ably assured. Where a program combines staff with concrete components such as sets of carefully structured materials, variations from one setting are likely to exist: one teacher's style may differ from another's, personality may enter into program use, and there are likely to be differences in the users' skills. Where a program lacks any concrete and tangible parts, it is still more subject to variation than those noted above. As what a program is built around moves from the concrete alone to include people, and finally to consist largely of people, the difficulty of establishing and maintaining control and hence program uniformity increases.

Other aspects of a program over which control is critical, if uniformity is to be maintained or diversity encouraged, are: (a) what is done in the program and (b) program goals (Weiss, 1972b; Stake, 1969; Cohen, 1970). First, what is done in the program – included here are the various components or stages in the program, their sequence, and the procedures to be used. As noted earlier, where any or all of these are set, uniformity of program is likely. Where users are able to select from among alternatives, limited variation will ensue. Where users are invited to develop their own components or sequence or procedures, the form of the program will vary from one site to another.

Another aspect of what is done in a program lies in the methods, techniques, and procedures used. Where these are set, it is likely that, within limits, the same techniques will be used for the program despite different locations. Where techniques are not specified or where users are expected to add or develop their own procedures, variation is assured.

Second, a program's goals may be set out so that users are expected to achieve all of them. Users may also be able to select from the total set of goals those to which they wish to give special emphasis. They may even be encouraged to devise their own goals for use with materials supplied by the program. Once users are able to devise their own goals, the program will vary from one group of users to another.

Further, where goals are stated in detailed and specific terms, there is little room for program users to interpret them. Where goals are broadly defined, users may develop particular emphasis and interpretations. There is more latitude for individual variation irrespective of whether or not the users are expected to achieve all goals. Again, monitoring by a central agency may serve to check such variations.

Other aspects of a program could be taken into account in considering uniformity. A program's size, length, location, and dispersion may, under certain conditions, affect uniformity but each will not necessarily

do so (Cohen, 1970). For example, a program located in a single classroom, which presents teachers with only an outline of possible activities, would show considerable variation from day to day.

If the program operated in two different locations, there could well be little similarity between them. A program which consists of installing a new kind of electric warning device at level crossings throughout the country in the hope of reducing accidents, though much more scattered and widespread than the first, would exhibit fewer variations. Yet size, length, location and dispersion may influence program uniformity (Weiss, 1972b). It seems likely that, as a program extends from a few hours through several weeks to years, variations in its use are likely to occur. Similarly, when a program operates at sites located in twenty different States, variation is more likely to occur than when it is located at two sites in the same area. It is more difficult to control a program that is widespread than one that is operating only at program headquarters. Again, if one program has a large target population and another deals with only a few persons, variation is more likely to occur in the former than the latter. To deliver identical service to 100 000 is more difficult than to do so for twenty. Since such factors are less critical in their effect on the program's uniformity than those discussed earlier, they are not included in the summary below and subsequent discussion of the case studies.

To sum up, aspects of a program likely to affect its uniformity include: (a) what makes up the program or what is used in it (materials or things, the people conducting the program), (b) what is done (the components or stages included, the sequence of activities or treatments, procedures or methods used), and (c) the program's goals (narrow, specific and particular, broad and complex, none given). Each of these aspects, and the total program, may be controlled and hence made uniform and standardized by certain procedures: in (a), by supplying materials and staff, or training local staff to carry out the program; in (b), by setting as compulsory the components to be used, the sequence in which they must be used, and the procedures for using them; and in (c), by the kind of goal used in the program and by making stated goals compulsory. Uniformity may be further enhanced by the process of monitoring the program by a central agency. (Too often what program directors assume is taking place in the classroom or on the site is not what is actually happening.) As program directors do not monitor the program and do give users opportunity to select among alternative components, procedures, and sequences, or even to devise their own, the program's form will become varied and diverse.

A summary of the characteristics is set out below: (In each cluster, items move from those likely to assure program uniformity to those which would contribute to program diversity.)

A What is used in the program:
- 1 established as:
  - (a) materials – supplied
    specified
    one's own devising
  - (b) staff – supplied
    trained
    untrained
- 2 monitored by a central agency

B What is done in the program (the activities and treatment):
- 1 established as
  - (a) the program components or stages – set
    selected
    devised
  - (b) sequence – set
    selected
    devised
  - (c) procedures – set
    selected
    devised
- 2 monitored by a central agency

C Program goals:
- 1 established as
  - (a) narrow – set
    selected
    devised
  - (b) broad and complex – set
    selected
    devised
  - (c) no goals given
- 2 monitored by a central agency

Not every program will include each of the elements set out above. Yet the elements given may be combined to give some indication of the uniformity of a particular program. Some possible combinations are given below, moving from a highly standardized uniform program to one in which the emphasis of program directors is on diversity and adaptation to local conditions.

Where a program uses set materials, to be worked through in a set order, according to specified procedures, to accomplish narrowly defined goals, the program is highly uniform. Its uniformity may be maintained by a central monitoring agency. Where program directors specify clear goals, supply users with materials and training, and permit selection among suggested procedures and stages, the program will exhibit some limited variation from one site to another. Where program directors leave goals broadly defined and, while suggesting possible materials, sequences of activities, and procedures, encourage local users to devise their own, the program will exhibit considerable diversity from one site to another.

Characteristic IV: Specificity of Program Objectives

A goal constitutes something to aim at. Goals vary in precision and complexity. Where they are broadly stated, it may be difficult to ascertain what the goal is, what would constitute fulfilment, and when it would be accomplished. If the goal is narrowly specified to refer to particular activities accomplished within a defined time frame, the end being pursued becomes clear.

First, goals vary in the way they are stated. Take a program whose directors wish 'to enhance the quality of life'. The goal is a broad one. Each of the key terms in the phrase, 'quality' and 'life', suggest many possible ideas, concepts, and activities. Some of these, no doubt, would conflict or be inconsistent with others. Weiss (1972b, p.5) describes the dilemma:

> A goal such as 'improving the quality of urban life' contains within it not only a large number of subgoals (that must be made explicit) but also ambiguous subgoals (improving the esthetics of the urban scene) that pose awesome problems of conceptualization and measurement.

In the process of seeking, what is sought may be altered. Initially program directors could recognize quality of life as related to leisure activities. Subsequently they could include quality of services, pollution-free environment, security, and so on. There are programs with goals that are still more open than those stated in broad terms. some programs set goals which are not intended to be fulfilled, reached, or achieved once and for all. For example in the arts and humanities, goals may aim at the student's taste, judgment, and critical sense. The goals are not intended to be fulfilled once and for all, but to give the student a chance to respond individually. The end is left to the participant's own devising (MacDonald-Ross, 1973, pp.28–29). Alternatively, take a goal like 'to visit the zoo'. The children are to visit

the zoo not in order to achieve some predetermined response or outcome, but to make their own expressive response in whatever form they wish (Eisner, 1969). Their response cannot be detailed in advance of the visit; that does not mean it cannot be assessed subsequently. The difficulty with such goals, unlike that with broad abstractions, lies not in possible conflicting interpretations – 'to visit a zoo' is clear enough – instead, it lies in the goal itself. The end point is not determined in advance. Hence the goal is open-ended. In contrast, program goals stated in behavioural terms are an extreme effort to obtain clear-cut end points for achievement. Mager's formulation (1962) attempted to clarify goals by insisting that they be stated in terms of observable actions. Efforts to narrow objectives need not arrive at the kind of behaviour definition recommended by Mager (1962). The goal, while clear, may refer to a pattern of behaviour or a state. For example, when a program goal is phrased like this,

> By the end of the course, students will be able to prepare and care for clay and clay objects from construction through the final firing,

what is being aimed at is clearly identified and reasonably narrow in scope (MacDonald-Ross, 1973).

Second, goals vary in the extent to which the area in which they are to be accomplished is specified. Sometimes the area in which a goal might be sought is almost limitless. In the quality-of-life example, given above, so much could be encompassed by the term that it is difficult to know any area of life in which the goal could not be sought. In contrast, the study of 'living organisms' or the 'mathematical concept of limits' gives a more precise area. In the ceramics goal above, the field is defined in reasonably clear terms.

Third, what constitutes acceptable fulfilment of the goal may vary in the clarity with which it is stated. In the 'quality-of-life' goal above, it is not clear what would be accepted as enhancing life. Would people be happier than at present? Have cleaner air than now? Undergo a fundamental change in the priorities assigned to public and private transport? In contrast, the goal for the ceramics program, given above, does give some indication of what would be accepted as fulfilling the goal. While not explicitly stated, students know that they should look for the various steps between preparing clay and firing it. Once familiar with those, it would be reasonable to assume that if a student could not wedge clay, he would be unable to meet the goal set out in the statement. If not only the performance desired but also the level of acceptability is defined in the goal, what constitutes fulfilment is tightly specified. The level of achievement may be made explicit by giving, for example, the percentage of those doing the ceramics course who must attain a certain level.

Fourth, goals vary in their time frame. The time in which a goal is to be accomplished may be specified, it may be unclear, or it may be left to the participants to set for themselves. Further, the achievement of the goals may be set for some period during the program, at the end of the program, or at some future time. For example, the time period for 'enhancing the quality of life' is left unclear. As achievement of the goal lies further in the future, fulfilment becomes increasingly difficult to relate to the program. Given that program directors could identify what they were aiming at, it would still be difficult to establish the contribution of the program in improving the quality of life. On the other hand, the ceramics program goal identifies the time within which the goal is to be achieved – by the end of the course.

To sum up – first, the program goals or objectives vary considerably in scope. They may be directed towards some specified end. They may be complex, permitting variation in fulfilment; or they may be broad abstractions, not only subsuming many complex subgoals but also some that may be incompatible. Goals may even be open-ended. Second, program goals vary in delimiting the area in which the goal is sought. Where the activity and conditions are explicitly stated, the area is tightly specified. But the area may also be indicated by a range of activities and states, or left largely unspecified. Third, program goals vary in the degree to which achievement of the goal is specified. It is most narrowly specified when not only the desired performance but the level is stated. Yet some program directors may be content to indicate the performance alone or even to leave it largely unspecified. Other program directors may leave the task of determining acceptable fulfilment of the goal to the participants. Fourth, program goals vary in their time frame. This may be specific, largely undetermined, or self-selected. Goals may be accomplished during the program, at the end, or may be designated to be achieved after the program is completed. Each of the four elements identified contributes to the specificity of program goals. They are set out in summary form below (with the elements which are the most precise and closed first, then moving to those that are open-ended):

A Statement of the goal:
  1 narrowly specified
  2 complex behaviour or states
  3 broad abstractions
  4 open-ended

B Area in which the goal is to be sought:
  1 conditions as well as activities are stated

2 range of activities or field is indicated
3 area is indeterminate, i.e. lacks clear parameters

C Specificity of acceptable performance, i.e. what constitutes fulfilment of the goal:
1 performance and level are specified
2 performance is indicated
3 performance is largely unstated
4 performance is selected by program participants, staff, trainees, or clients

D Specificity of the time frame for achievement of the goal:
1 specified
indeterminate
self-selected
2 during the program
at the end of the program
long term

These may be combined to give some indication of the treatment of objectives in a particular program. For example, if program directors specify the time by which goals are to be achieved in the immediate future (e.g. by the end of the unit, the course, or the program), state the conditions under which the goal is to be sought, the level of acceptable performance, and define the goal in terms of an observable behaviour, the program goals may be said to be precise terminal outcomes.

If program directors set complex behaviours in some broadly defined field and leave unstated what would be acceptable performance, program goals are still reasonably clear. If program directors set open goals, they expect participants to use the program in a variety of ways in accordance with their own purposes. Responses cannot be determined in advance, and what constitutes an acceptable outcome is for participants, not directors, to decide. In effect, the program goals are not specified.

CHARACTERISTICS RELATED TO THE EVALUATOR'S STANCE

Characteristic V: Evaluator's Independence

Two aspects of independence are important in an evaluation study: the evaluator's independence of certain groups or persons, and the evaluator's independence to do what he or she sees fit in conducting the study. The first, independence of, points to *who* may control the

evaluator, that is, to the extent of the evaluator's autonomy. The second, independence to, points to the evaluator's freedom of action, that is, to what is done during the evaluation study. The two are closely related and mutually dependent. If one is independent of someone, one is able to follow a selected course of action without interference. Conversely, if one is dependent on someone, that person is able to influence one's actions. The evaluator's independence of others is synonymous with the extent to which he is free of their control and is not subject to their authority or under their jurisdiction. The issue is: *Who*, in the context of an evaluation, is able to impose *what* sanctions or influences on the evaluator, thus limiting his or her autonomy?

Two sanctions are critical: the control of funds for the study, and the power to hire and fire the evaluator (Glock, 1961). If an evaluator's funds are withheld, or he is fired, the study ceases. Either sanction may be imposed at any time in the course of the evaluation. Imposition of either sanction affects the evaluator's future. Thus those who fund the evaluation and hire the evaluator have much power over him. Further, if neither sanction is exercised, that does not indicate that they do not exist, only that their inherent power remains latent. This power may, for example, be used to influence the shape of the study and, more particularly, what is reported. Funding agencies may well have ideas as to what should be included in an evaluation as well as the most appropriate procedures for the evaluator to use. It is here that the close connection between the evaluator's 'independence of' and 'independence to' becomes apparent.

Program directors and project personnel may exercise other controls less critical than funding and fate control (Bend, 1970). They frequently are in a position to influence directly what the evaluator does. They may prevent access to certain information; they may limit the evaluator's potential use of experimental and control groups; and, more generally, they may inhibit the flow of information from program to evaluator. If they do so, they exercise considerable control over the evaluator's independence to conduct the study as he or she sees fit. Further, their influence is likely to be felt during the evaluation, affecting design and data collection as well as content and form of reports. Similarly, those to whom the evaluator sends reports may influence what is included in them or, more subtly, the treatment of the information presented.

Yet those whose *only* connection with the evaluator is through reports are in a less powerful position to exercise direct influence over the study than other groups noted above. Groups that only receive reports have no direct control over the evaluator's fate. Reports typically reach

them *after* the study has been completed. Criticisms at that stage may affect what an evaluator does in subsequent studies, and possibly his subsequent career opportunities. They cannot alter what has already been done, although the evaluator's anticipation of the group's interests may have shaped his study. A similar influence on the evaluator's work may be exercised by those whom the program serves: for example, the community, students, administrators, etc.

To summarize, the evaluator's independence may be limited by any one of several groups: the funding agency, the hiring agency, program personnel, and those receiving reports and those served by the program. In a particular evaluation, each of the sanctioning or influencing powers may be held by a separate group – that is the way they have so far been treated – but these powers may be combined so as to be under the control of one or two groups. Many different combinations are possible but the critical ones relate to the evaluator's independence of program personnel.

First, consider a situation in which an evaluator deals with a single group. Here the group developing the program that is to be evaluated has funding and fate control over the evaluator, and is the sole recipient of evaluation reports. The evaluator thus has little space in which to exercise autonomy. If program personnel wish to influence the evaluator's actions, they have several avenues open to them. While the evaluator may negotiate a measure of independence, he has no court of appeal. His autonomy is minimal.

When authority is dispersed among two or more groups, the evaluator's autonomy may be enhanced, yet much depends on which sanctions particular groups control. If program developers hire the evaluator and fund his work, while there is a second group to whom the evaluator also reports, the evaluator's effective autonomy is increased little. However, once either the power to hire the evaluator or the control of evaluation funds is out of the hands of program personnel, the evaluator does have some measure of autonomy. If the evaluator is to be fired or funds stopped, agreement between two parties is necessary. The evaluator has a court of appeal.

The evaluator's independence is further strengthened when a group other than those conducting the program hire him and fund his study. Here program developers may still affect the flow of information to the evaluator, but they are no longer in a position to impose critical sanctions.

Again, control may be dispersed among three or more groups. Potential users of a program may secure funds for an evaluation from an outside source, thus dispersing control between program developers, funding agency, and the users themselves. Here the evaluator has to

deal with several groups, but possible sanctions and influences over his work are widely dispersed. Decisions to withdraw funds from the study would involve co-operation between several groups. The evaluator's autonomy is reasonably secure – although his task may not be an easy one.

Finally, there is the situation in which the whole project is under the evaluator's control. Here the evaluator controls the funding, possibly even that of the program being evaluated. He would be in a position to control fully the design of the study, data collection, and reports, and to shape the program to the needs of the evaluation. Besides the relationships among groups holding sanctioning power, two other factors affect the evaluator's autonomy of action: the evaluator's place in the organizational hierarchy, and the extent to which the evaluator is insulated from demands of program personnel.

An evaluator, who is not a 'free lance', has a subordinate role in the program's organizational hierarchy. There are people to whom the evaluator is expected to report, and with whom he interacts: these may influence what the evaluator does. The organization may not necessarily be a single unit. Since the evaluator's role demands that he deal with several groups – which may belong to different organizational units – he or she is a member of a bureaucratic organization including those various units (Weiss, 1972b, p.21). The position held by the people to whom the evaluator reports in turn defines the level of the evaluator's authority in the overall bureaucratic organization.

Thus, if an evaluator is in a position to report directly to the directors of the program or the funding agency, he has a reasonably high status within the organization. If, on the other hand, the evaluator has access to the funding agency only through the program directors, or if he or she has access to the directors and funding agency only through their administrative assistants, then his position is a subordinate one. When an evaluator has to report via others, they are in a position to select what will be presented and even to block evaluation reports. In effect, intermediaries who stand between the evaluator and his ultimate audience act as successive filters on the study's reports. Further, when the evaluator is a junior member of the organization, he is not in a strong position for resisting the influence of program personnel (Gouldner, 1965). Related to the evaluator's status is the location of the evaluation unit. When an evaluator is part of the program's organization, he is likely to be physically located in the same block of offices, and thus will have little chance of insulating himself from program demands. Yet he may be able to build in some checks and so give himself some measure of autonomy. An evaluator's autonomy may be increased by a location which, while enabling him

to visit the program site, leaves the evaluation unit functioning separately from the program's organization. He may even conduct the evaluation at a location remote from program headquarters, and collect data by mail. Here the evaluator is well insulated from the influence of program developers or users. (On the other hand, the lack of contact may affect the evaluation study. The evaluator who operates by remote control may not understand the program as fully as one who has some contact with it.)

To sum up, in considering the evaluator's independence, the factors to be taken into account include who controls the sanctions and the evaluator's place in the organizational hierarchy. The greatest independence is when the evaluator funds the program and constitutes the major audience for the reports. Here the sanctions are controlled by the evaluator. Where the sanctions (hiring, funding, and reporting) are shared among several groups, the evaluator's independence is less. Several arrangements are possible: the program developers may receive evaluation reports, but the evaluator is funded and hired by another group who are also recipients of the evaluation reports; the program developers hire the evaluator and receive his reports, but the evaluator is funded by another group to whom he also reports; and the program developers hire the evaluator, fund his study, and receive his reports, but there is also another group to whom the evaluator reports. The least independence is where the sanctions are controlled by a single group: the program developers hire the evaluator, fund him, and are the sole recipients of his reports.

The evaluator's place in the organizational hierarchy relates to his access to key project personnel. He may have immediate access, and report directly to key persons, for example program directors, funding agency directors, etc. He may hold a subordinate position and report to the funding agency via program directors or to program directors via administrative assistants. He may even be a junior member, with his reports and requests filtered through various levels of the organization.

The position or location of the evaluation unit in relation to the program unit also affects the evaluator's independence. The evaluation unit may be remote from the program organization; here the evaluator has no face-to-face contact with the program users or operators and all evaluation data are collected by mail. The evaluation unit may still be separate, but the evaluator visits the program sites and receives direction and guidance from program personnel when he solicits it. Alternatively the evaluation unit may be part of the program organization, located on site, yet with checks on the influence program personnel may have on the evaluation. The greatest influence is where the evaluation unit is part of the program organization – located on site;

and there are no checks on the influences, overt and covert, which program personnel exert on evaluation personnel. The elements of the characteristic are summarized below:

A Who controls the sanctions:
  1 the evaluator, who also funds the program
  2 shared by several groups, for example:
    (a) funding and hiring separate from the program developers
    (b) program developers hire the evaluator, but the funding is separate
    (c) program developers hire and fund the evaluator, but the evaluation reports are also sent elsewhere (e.g. project funding agency)
  3 program developers fund and hire the evaluator and are the sole recipients of the evaluation reports

B The evaluator's place in the hierarchy:
  1 immediate access to key project and funding personnel; has a senior level position
  2 limited access to key project and funding personnel; has a subordinate position
  3 reports to key project and funding personnel, filtered through levels of the organization; has junior position

C Position of the evaluation unit:
  1 remote from the program organization
  2 separate from the program organization
  3 part of the program organization
    (a) with insulation
    (b) without insulation

The elements set out above may be combined to show the evaluator's independence in a particular study. Four examples are given below.

Where those conducting the evaluation control the funds (their own funds and those of the program) and constitute the major audience for reports, the evaluator's autonomy is extensive. There is no group with effective sanctioning power over the evaluator.

When the evaluator holds a senior position in the program's organization, and is neither hired nor funded by program directors, the evaluator is in a position of considerable autonomy.

When the evaluator has been hired by program directors, holds a subordinate position in their organization, but controls a separate evaluation unit and independent funds, he still has some measure of autonomy.

When program directors have hired the evaluator and control the

funding for his work, and the evaluator is a junior member of their organization, he has little space in which to exercise autonomy. His independence is minimal.

Characteristic VI: Evaluator's Orientation to Value Judgments on Program Merits

'Value judgments are judgments of value, worth, or merit.' (Scriven, 1974) Some are simple matters of taste, as when someone expresses a preference for green over red. Others are more complex than this. Where the context is defined, the value judgments may still be clearly identified and checked. But commonly, values express views on the relative merit of complex matters. The 'democratic way of life' or evaluation may be valued primarily as a tool in curriculum development. There may well be different opinions as to what should count as good (Scriven, 1967, pp.69–70). In the context of an evaluation, the value judgments which are critical are those that are concerned with the worth or merit of the program.

Value judgments may be unacknowledged in a study but still enter into it. They may be casually noted, as when an evaluator comments that a certain group of users judged the program as valuable. On the other hand, value judgments may be deliberately addressed in the study. An evaluator may describe his own frame of reference as well as the standards he has used in assessing a program. An evaluator may even follow Stake's (1967) advice and record opinions of those connected with the program, accepting the responsibility of portraying the merit and fault of the program as 'perceived by well-identified groups'. He may make it part of his task to ensure that the 'scope and diversity of opinions of the community are known' (Stake & Denny, 1969, p.388). There are two obvious sources for opinions about the merit of a program: the evaluator and program directors. But there may be others whose value judgments are relevant to the program: those using it, its clients, and their families. For an educational program, the list may be extended to include subject-matter experts, teachers, parents, students, and those able to speak for society in general (Taylor & Maguire, 1966).

The existence of several groups whose value judgments may be relevant to the program opens up several alternative courses of action for the evaluator. He may confine treatment of values to his own, using them alone to assess the program and guide the study. Or he may seek out the judgments on the program from groups affected by it, conducting his study in a way that takes account of their different value systems. Or he may present the value judgments of a particular group or person other than himself – typically, program

developers. In doing so, possible differences in judgment on the merits of the program are omitted.

Finally, there is the way the value judgments are presented in the study. An evaluator may not only collect and present information on a program, point out its strengths and weaknesses, but also deliver judgment on the program's merits: 'Overall, the program must be judged ineffective,' 'The program has exceptional merit,' and so on. The verdict is reached and presented irrespective of the feelings of program personnel or sponsors, and irrespective of possible consequences to the program. In whatever language it is couched, the verdict on the program – and there may be several such judgments – is not left to someone else but is presented by the evaluator as part of his own work.

An evaluator may leave the judgment on the overall merit of the program to others, limiting the use of value judgments to suggesting strengths and weaknesses in the program. On occasion, information about a program may be presented and used as a basis for posing questions and raising issues, without actually stating the program's merits. Or he may limit the task to presenting information about the program. He may describe what went on in the program, what data were collected, and how they were analysed, lay out the results of the analysis, and then stop. The reader, not the evaluator, is thus faced with the task of judging (Worthen & Sanders, 1973).

To sum up, in considering the evaluator's orientation to value judgments, relevant factors include whose judgments are used. These may be the evaluator's, those of several groups including the evaluator, or those of the program developers alone. Value judgments may be left largely unacknowledged, casually noted, or given explicit recognition in the study. The degree to which the evaluator is willing to make judgments varies: some evaluators may be prepared to judge in the sense of reaching and making explicit their verdict on the program and its several aspects; others may judge by listing the strengths and weaknesses as they perceive them; or the evaluator may be content simply to lay out the data and leave the judgment to others. The elements of the characteristic are summarized below:

A Visibility of value judgments in the study:
  1 unacknowledged
  2 casually noted
  3 systematically set out

B Whose value judgments are included:
  1 the evaluator's alone
  2 those of several groups or persons

3 the program developer's alone, or those of some other interested party

C The way value judgments are used:
1 not used – the data or 'facts' are simply set out
2 to suggest strengths and weaknesses in the program
3 to give a verdict on the merits of the program

The elements set out above may be combined to indicate the way a particular evaluator has treated judgments on the merits of the program in his study. Some possible combinations are sketched below, moving from those where judgments on the merits of the program are minimized to others where they play an important role in the evaluation.

An evaluator may pay no explicit attention to value judgments on the program's merits in the course of his study. Thus, while his own values function implicitly as the frame within which the evaluation is conducted and the program assessed, explicit attention is confined to setting out the 'facts'.

An evaluator may set out explicitly the values and standards he has used in assessing the program's merits. He may use these to raise questions about the program, to point out possible strengths and weaknesses, leaving the task of reaching a verdict on the program's merits to others.

An evaluator may also include in the information gathered, the judgments of groups affected by the program. Reports may set out where the groups differ in their assessment of program merits and use these as a basis for reaching a final verdict on the merits of the program. Here the role of value judgments in the study is extensive and open, with alternative views presented. The task of judgment is included in the study and not left to others.

Characteristic VII: Evaluator's Orientation to Utilization of His Work

Decisions are said to be 'informed' or 'uninformed', depending on whether they are based on information or made in ignorance of relevant information. Decision-makers use information, evaluators collect it. It is through its information-gathering function that an evaluation study may be *used* by a decision-maker. To make a decision is to use information to choose between alternative courses of action. The use of information in decision-making thus covers a cluster of factors which, once identified, serve to further define and clarify the characteristic. Let us now look at these factors.

First, if an evaluator expects his study to be used, he needs to *identify likely users* in the course of the study – preferably near the beginning of the evaluation process (Ferman, 1969). The range of possible users of evaluative information is extremely wide and without clear bounds. Obvious candidates include those who developed the program and those that funded either the program or the evaluation. Less obvious ones include future developers of similar programs, evaluators confronted with similar programs, and the users or clients of the program. As the circle of possible users widens, tracing how, when, and where the evaluative information is used becomes difficult and tenuous. But that need not cause trouble as the issue here is simpler: did the evaluator identify likely users of information gathered in the course of the evaluation study? The evaluator, having identified several such users, may choose to serve only one to two or may attempt to serve them all.

Second, a decision is a choice between alternatives. The number of alternatives that are laid out for the decision-maker is thus important. A single choice (to continue/discontinue the program) may be the focus of the evaluation, or the evaluation reports can lay out alternative courses of action indicating the benefits and disadvantages associated with each. A single recommendation (to continue/discontinue) requires data on the program as a unit; but, if several alternatives (to modify this or that aspect of the program) are to be laid out, data need to be collected on different components of the program (Glock, 1961).

Embedded in the choice between alternative courses of action, is the issue of the *way in which the evaluator presents the data collected*, that is,. the extent to which the evaluator makes ready the information for use by others. The data may be presented and left to speak for themselves; interpretations may be made of the findings; alternative courses of action may be suggested; and, in addition, the relative strengths and weaknesses of each of the possible courses of action may be assessed. In preparing information collected for use, the evaluator is, to some extent, involving himself in the decision-making process. He may do so actively and directly by personally advocating his recommendations (Longood & Simmel, 1962). An evaluator may also pursue the use of information from his study by including likely users in designing and conducting the evaluation (Rippey, 1972; Weiss, 1967), including formulating recommendations.

Third, if evaluation findings are to be used in making decisions, they need to be *available in time for inclusion.* Information, however relevant and sound, only available several months after the decisions have had to be made cannot be used in making those decisions. On

the other hand, information tabled prior to a decision deadline may be of use in drawing attention to the nature of certain decisions.

Fourth, if evaluative information is to be used, it needs to be *pertinent to that use*, that is, to the decision being made. Put another way, the information needs to be collected on issues of relevance to users. When the continuance of aides in a reading program is at issue, information on the performance of the pupils in the program is not directly pertinent. When a program is faced with budget cuts, information that deals only with the program as a whole is of little use in deciding which aspects should be curtailed.

Fifth, if the evaluation study presents information both timely and pertinent to the decisions being made but in a form largely unintelligible or opaque to users, the latter will have difficulty in basing their decisions on that information. Lay persons, faced with a report largely couched in some species of technical jargon or using multivariate analysis without adequate explanation, are usually not in a position to make a decision based on that information.

Sixth, the likelihood of an evaluation being timely and pertinent and laying out alternative courses of action may be increased where the evaluator has *access to the policy makers* at higher levels than the program directors, or when there are key outsiders who will bring pressure to bear if evaluation results are ignored (Weiss, 1967). The list could be extended, pointing to other organizational structure features of the evaluation which affect the use by decision-makers.

To sum up, the factors identified can be clustered around two aspects of utilization: who and what. The first includes the identification of potential users, the involvement of the decision makers in the evaluation, and the timing of the reports. The second includes the identification of the decisions which may be made, the pertinence of the information collected to those decisions, and the presentation of the information. The characteristic is summarized below:

A Who:
- 1 identification of likely users
  - (a) no specific user identified
  - (b) one major user identified
  - (c) several potential users identified
- 2 the involvement of the decision-makers in the evaluation
  - (a) they are removed from the evaluation process
  - (b) they are involved in the planning of the evaluation
  - (c) they are participants in the evaluation
- 3 timing of the reports
  - (a) evaluator's design determines the report timing
  - (b) the decision cycle determines the report timing

B What:
- 1 identification of the decisions to be made
  - (a) none specifically identified
  - (b) a single global decision identified
  - (c) several alternative decisions identified
- 2 pertinence of information collected
  - (a) of general interest, not specifically related to the decisions
  - (b) pertinent to the decisions
- 3 presentation of the information
  - (a) intelligible, but not suggesting action
  - (b) in a format which suggests possible courses of action
  - (c) advocacy of recommended action

A single evaluator may not direct his work towards users in each of these ways. Yet it does seem reasonable to conclude that the more of the aspects addressed in the evaluation, the more the study can be said to be oriented towards utilization and decision-making; the fewer the aspects that occur in an evaluation study, the less it is oriented towards decision-making and utilization of its findings.

The elements in the characteristic listed above may be combined to indicate the pattern of orientation towards decision-making/utilization in any particular study.

Where the evaluator makes no clear identification of possible or likely users, and decisions follow the demands of his selected design in timing the presentation of reports and in their format, the evaluation is little oriented towards utilization. Where an evaluator does identify likely users of the study and decisions, collects information partly in terms of those decisions, and in reporting on the study does indicate some relationship between the information collected and likely decisions, the evaluation study may be said to be partially oriented towards utilization. Where an evaluator, in addition to identifying likely users and the uses to which the study will be put, includes the users in planning the evaluation study, and delivers the information in time for it to be used in making decisions and in a form which is not only intelligible but also lays out alternative courses of action suggested by the data, the evaluation study is clearly oriented towards utilization.

CHARACTERISTICS RELATED TO QUESTIONS ASKED IN THE EVALUATION

Characteristic VIII: Attention to Program Goals

There is a necessary relationship between an evaluation and what is

evaluated. In an evaluation, something has to be evaluated; the program is central to the conduct of the evaluation. The relationship between program and evaluation varies with the extent to which the evaluator conducts the study within terms defined by the program.

Programs have many explicitly defined aspects – goals, processes, events, structures, and so on – but frequently the aspect that is stressed is the program goals or 'intended outcomes'. Each defined aspect of a program represents a selection from a universe of possibilities. The evaluator, faced with the program's selection, may limit his study to aspects defined by program directors. In this case, the intended outcomes specified by program directors largely determine information the evaluator collects. However, programs usually accomplish more than what is specified by their directors. In addition to the explicitly stated goals, there are implicit ones; other effects than those sought are present; other procedures than those specified are used; and so on. Each program has a range of concomitant features. The term 'concomitant' is used in preference to the more usual ones of 'hidden' or 'side effects' or 'unanticipated' outcomes, as each of these carries predominantly negative connotations. As Scriven (1974) has pointed out, the features are not necessarily negative. In conducting an evaluation study, the evaluator who pays attention to concomitant effects of the program may seek only information on harmful effects and ignore the possibilities of positive spin-offs. Or he may limit attention to additional positive program outcomes. Or he may collect data on several aspects of the program (e.g. events, procedures). Further, the extent to which studying the concomitant effects of a program is central to the evaluation study may vary. For example, in conducting the evaluation, concomitant effects may be picked up incidentally in the course of other data collection, or they may be part of the formal data collection (and anywhere along the continuum in between).

Here a distinction must be made. The situation described above refers to an evaluator working from *within* a program. An evaluator may also go *outside* a program's specified aspects. Questions may be asked as to what other events or procedures or activities or sequences could achieve the same goals. Or, how important is the program, that is, are there other viable alternatives already available to the population the program intends to serve? How large is that population? And so on. In taking such a position, the evaluator puts himself outside the mandate of the program and in a position to question every aspect of it, including the most basic aspirations of its developers. To sum up, in conducting an evaluation study, an evaluator may treat the program in three ways: the evaluation may be limited to collecting

data on goals and features of the program that are specified by the mandate of the program; the evaluator, although limiting the examination of the program from within, may include data on unspecified outcomes and features of the program; or the evaluator may study the program from without, the data being collected on the importance of the program in the broader societal and educational context. The characteristic is summarized below:

The evaluator treats
A the program goals as:
  1 incidental
  2 integral

B the concomitant aspects including:
  1 outcomes
    (a) incidental
    (b) integral
  2 processes
    (a) incidental
    (b) integral

C the aspects outside the program as:
  1 incidental
  2 integral

The elements listed above may be combined in several ways to give some overall indication of the attention to program goals in a particular evaluation. Where the evaluator conducts the study in terms of the program, concentrating on its goals, features, and processes and seeking answers to questions determined by program directors, the evaluation is conducted within the mandate of the program. An evaluator may also attend to matters not included in the goals of the program. Minimally, an evaluator may incidentally gather information on one or two unintended and undesirable outcomes. He may, in addition, incidentally collect a little information on the program's place in a broad social and educational context. In either case, the evaluator, while concentrating evaluative effort on issues specified by the program directors, is treating additional issues incidentally.

However, an evaluator may also give attention to concomitant program effects integral to the study at the same time, and assess its importance in its social and educational context. If an evaluator does so, factors other than those specified by program goals make a solid contribution to the study.

Characteristic IX: Generalizability

A particular event is isolated in several ways from other events. It occurs at a particular time and in a particular place, and what is instanced in the event is, itself, unique. Our term, unique, signifies something so outstanding as to be the only one of its class. The unique is particular. It may have a high degree of generality but it cannot, taken on its own, be generalized from, that is, taken as giving information about other events. Yet a single event may be treated as a *basis* for generalizing to a larger set of events. The event may be related to others in several ways: across time (earlier or later instances), across space (i.e. the people in the next town, state, or country), and to a number of specific instances of the general phenomenon (i.e. other instances of the same event). Further, even if an event is able to be generalized in all three ways (time, space, and instance), the extent to which the generalization is carried may vary. The event may have only limited applicability to other specific instances of the same program. It may afford the possibility of generalization to programs of a similar type or, if it is a program that embodies a well-defined construct from research, it may afford the possibility of extensive generalization, that is, of generalizability.

There are several indicators of whether an evaluator is treating the program as an isolated event, enabling generalizations to be drawn from it. There are the questions that an evaluator sets out to answer, the way the information is sought, and the discussion of what has been found. Each is briefly outlined below.

The questions to which an evaluator directs the study may be program specific. Even where program inputs, processes, and outcomes are generalizable, the evaluator may treat them as pertaining only to a particular program, so ruling out the possibility of a more generalizable evaluation study. For example, a program may have as a process some form of the discovery approach. The evaluator, in examining the program, limits attention to the particular instance presented in the program, not relating it to 'discovery' approaches in other programs or to information collected in other evaluation studies.

There are, however, many ways in which an evaluator may extend the questions addressed in his study beyond the limits of the particular program being evaluated. An evaluator may treat a program using Computer Assisted Instruction (CAI) in terms of other research and evaluation studies that have used CAI. The particular program evaluated is treated as an instance of a type of program. Possible generalizations would be limited to other programs of a similar type.

The questions to which answers are sought by an evaluator may have a still wider import, dealing with the constructs and theories of basic research. For example, an evaluator studying a program that claims improved student motivation may choose to examine, in the evaluation, theories of motivation advanced by research workers. The discussion of what is found in the course of an evaluation is the other side of the coin to the questions asked. Thus, like questions asked, discussion may treat the program in isolation as a particular instance, as an instance of a type, or in terms of the theoretical constructs developed in research.

To turn now to the relationship between possible generalizability of an evaluation and the way the information is gathered in the evaluation, standard educational research procedure indicates that, if information is to be generalized, the drawing of the sample and the instruments used in gathering information are critical.

First, the sample: A sample, if drawn correctly, permits inferences to be drawn from the particular situation in which the data are gathered to a wider population. The evaluator may thus generalize from the program under study. It must be noted that attention to sampling is not limited to experimental designs based on agrarian research models. Both case studies and participant observation designs, based on carefully drawn samples, are techniques that have been used to generate information that is generalizable (Smith & Pohland, 1974).

Second, the instruments used to collect information: If the instruments used in an evaluation are well established and documented, the evaluator will be in position to generalize to a wider context. In effect, the lines of generalization are already established and known; the instruments have been carefully documented. Techniques such as unobtrusive measures, participant observers, open-ended interviews, while offering a general format that may have been used elsewhere, are unlikely to occur in exactly the same form from one evaluation to another. They thus limit the extent to which the evaluator's results may be treated as generalizable. Even more limiting are the 'home-made' instruments (tests, scales, interview schedules, and so on) that are not tried out or systematically developed prior to use in the evaluation. While they may fit the program being evaluated more precisely than other techniques, they do so in a way that inhibits transference of the information gathered to another context. One further point – for each of the factors (questions asked, and their discussion, sampling, and instrumentation), there remains the issue of the proportion of the evaluation that is directed towards generalizable information. The evaluator may include only one or two questions

that treat the program in a way that permits generalizations. Similarly, information collected in only one part of the evaluation may be based on a systematically drawn sample.

To sum up, the generalizability of an evaluation is dependent on the questions to which the evaluator seeks answers, the way information is sought, and the discussion of the information gathered. The questions asked, and the discussion in the report, may treat the program as an isolated instance of one unique program, as an example of a type of program which may be repeated elsewhere, or an exemplar of theories and constructs developed in research. Sampling and instrumentation form two subdivisions of the way information is sought in the evaluation. The characteristic is summarized below:

A Questions treat the program as:
1 isolated instance
2 example of a type
3 example of a research construct

B The way the information is sought in the evaluation:
1 the sample
(a) limited to the program
(b) selected with possibility of inference to a wider population
2 the instrumentation
(a) well known and documented outside the bounds of the evaluation
(b) the general approach and principle are known, the details would vary with each use of it, e.g. participant observation
(c) the instruments are home-made, developed for and used in the evaluation study alone

C The discussion treats the program as:
1 an isolated instance
2 an example of a type
3 an instance of a research construct.

The elements in A and C are the same; if the evaluation is completely consistent, the two should correspond exactly.

The elements given above may be combined to indicate the extent to which a particular evaluation study is directed towards generalizability. Where the questions the evaluator asks are specific to the program, and the techniques used to gather information are largely home-made, there is little emphasis on generalizability in the study.

Where the evaluator's questions treat the program as an instance of a type, and use is made of systematic sampling and standardized instruments, there is some emphasis on generalizability in the study. Where the evaluator's questions and discussion include issues of basic research, and samples have been systematically drawn and data collected in a standardized way, there is considerable emphasis on generalizability in the study.

CHARACTERISTICS RELATED TO EVALUATION DESIGN, INFORMATION COLLECTION, AND INTERPRETATION

Characteristic X: Formality of Control

'Formal' may be used to refer to the stiff and starchy, the unbending and dry in conduct. It may also refer to conduct that proceeds according to accepted conventions or rules. It is in this latter sense that the term is relevant to the conduct of an evaluation. The use of rules or conventions that control a study increase its formality; these include the design chosen, the data collection procedures used, and procedures used for the data analysis. Where a study is highly formalized, the results are likely to be reproducible by independent teams and the conclusions advanced have a solid defence against alternative hypotheses.

As a group, experimental designs have an highly explicit formalized set of rules that tightly control the pattern of design and interpretation of results. In true experimental designs, formality of control is maximized and possible sources of error are minimized. Quasi-experimental designs (Campbell & Stanley, 1963) offer less control over possible sources of error than true experiments; yet, while the source of error or lack of control varies with the design used, in each case the source may be identified. The effect is to make the conventions or rules less clear-cut, to open up a wider range of alternative procedures and interpretations than in the case of the true experiment. Pre-experimental designs (Campbell & Stanley, 1963) give the evaluator so little formal control that Weiss terms them 'non-experimental' designs (Weiss, 1972b, p.73). Such designs include after-only study of program participants (Campbell and Stanley's 'one-shot case study'), either without or with a non-random comparison group ('static-group comparison'), and before-and-after study of a single program ('one group pre-test/post-test design'). Each of these designs leaves considerable room for varying interpretations of the extent to which any

change that is observed was related to the program and not to other factors.

There are, however, designs that are non-experimental in a different sense: they do not derive from the experimental frame of reference. Here these are called 'discursive' designs.* The total possible number of discursive designs is large, since broadly speaking they span all approaches used in the humanities and the 'soft' social sciences. They include, on the one hand, anthropological field work and, on the other, aesthetic criticism. Some approaches are shared by several different areas in the humanities. For example, 'structural' approaches have been developed and used in anthropological field work (Levi-Strauss, Frake, Conklin, for example) and literary criticism (Roland Barthes, Deridat, for example).

Phenomenology is another approach that shows a similar span of fields: Merleau-Ponty in literature, and Goffman in anthropology. Narrative approaches are shared by certain historians and critics, for example A. J. P. Taylor and Granville Barker, and Wilson Knight. In any one area of the humanities and social sciences, several approaches are likely to be in use. Literary criticism and history both include analytical, narrative and, recently, quantitative approaches. A similar range can be found in anthropology or sociology.

Within any one approach there is considerable variation. For example, Geertz's first chapter in *Interpretation of Cultures* identified aspects of his own field methods which are distinct from those of other anthropologists. Smith and Pohland (1974) similarly commented on the absence of a single unified approach to field work.

There is no clear-cut consensus regarding classification of discursive approaches. In a recent paper on 'qualitative knowing' in evaluation, Campbell (1974) distinguished the following approaches: site visits, systems analysis, field work, historical analysis, and adversary procedures of the legal system. His discussion suggests strong similarities between the first four. Recent discussions of evaluation would add criticism to Campbell's list (Mann, 1969; Weiss & Rein, 1970; Eisner, 1972). In addition there is also a long tradition in evaluation of using impressions (see for example, Judd's comments on it in 1918).

The classification in this study includes four discursive approaches: adversary proceedings, field work, criticism, and impressions. Variations within each of the discursive approaches make it difficult to identify systematically one as more or less formal than the others. The

*Campbell (1974) uses the term 'qualitative design'. To avoid confusion with the elements of data collection, and data analysis elements which are discussed later, the alternate term, 'discursive', has been adopted in this study.

rules or conventions governing each are largely left implicit or uncodified. Therefore, while the four discursive approaches are presented in order of formality of control, in specific instances the order may alter.

The adversary model generally lays out reasonably explicit procedures. Kourilsky's (1973) account of such a design requires both an affirmative and negative evaluator for each study. The case each evaluator prepares is partitioned into contentions logically related to the program and the value premises on which contentions are based. The submitted reports of the two evaluators are then cross-examined and defended, and the key decision-maker synthesizes views into a final decision. (Wolf's and Levine's accounts of adversary proceedings enumerate comparable steps. See Levine, 1973; Wolf, 1974.) While there is room for individual judgment, each step in the process is formalized. Evidence presented is subjected to systematic examination by some one other than the person who collected it.

Field-work models, including historical analysis and systems analysis, whether using direct participant observation or observation by a 'detached outsider', require continued sustained contact between observers and program. An evaluator using such an approach has to reside at the program site for extended periods (Campbell, 1974, p.19. See also Schulberg & Baker, 1968; Weiss & Rein, 1970). It is the close contact between the observers and project personnel which provides generally greater formality of control than criticism and impression, while it is the lack of formalized procedures which provides less formality than adversary proceedings.

In criticism models, the procedures and design are usually more implicit and unidentified than in field-work models. For example, field-work procedures often require that items be collected from two separate persons; that several methods of data collection be used for each item; and that items be checked in several situations and with different groups of people (Smith & Pohland, 1974). It is difficult to institute a similar list of counterchecks for collection of information and interpretations in criticism. The object of criticism is frequently either inanimate (art, literature) or, in the case of plays and concerts, ephemeral. Hence while like field work, in that it depends on close analysis of the phenomenon, criticism is more idiosyncratic.

A design that rests on collecting impressions from a single site visit provides the least formalized procedure. But, here too, the approaches vary. The design may simply be to spend a day on site, seeing as much as possible and then writing up any impressions. Alternatively the evaluator may collect informal impressions on more than one visit. In either case, the procedures lack the preparation of the cases for

and against of the adversary model or the sustained contact of field work and criticism.

Over all, in discursive designs, less formality is possible since the design does not constrain the paths that the evaluator may take to the extent of the experimental and quasi-experimental designs. Yet none of the approaches denies the existence of extensive prior planning by the evaluator, including the deliberate selection of a design. The difference lies in the extent to which the designs selected set out clear conventions or rules for their operation, thus controlling the lines along which the evaluation study may develop more or less firmly, that is, the difference lies in the formality of control.

There is no reason why in a single study an evaluator should not use both experimental and discursive approaches. He may use experimental design for testing outcomes, and participant observation for collecting data on the program in action. The degree of formality of control would vary with the designs incorporated into the evaluation. Associated with formality of design is the formality of procedures used in collecting data. Data collection procedures may be said to be formal or standardized to the extent to which data collected are reproducible; or to the extent to which the procedures remain the same, irrespective of who collects the data, from whom they are collected, when they are collected, where they are collected, and how they are processed. For each aspect to be standardized, it has to be explicitly defined so that ambiguity and possible variants are minimized. Possible standardized procedures include use of standardized measures such as achievement tests, IQ tests, and observation instruments or instruments developed and standardized by the evaluators themselves. Data may also be gathered by procedures that depend upon the qualities of the person collecting the information. These may include participant observation, personal reactions and experiences of those involved in the program being evaluated, and so on. Such procedures, while idiosyncratic in quality, are not necessarily haphazard. Participant observation, for example, includes systematic recording of the perceptions of others as well as the participant-observer-evaluator.

Finally, in an evaluation study some data may be collected by standardized procedures while others are not. The greater the use of standardized procedures, the greater the formality of control in data collection. A further aspect of an evaluation that contributes to formality of control is whether the data are treated quantitatively or through verbal analysis. Quantitative treatment includes inferential and descriptive statistics. Verbal analysis may include the construction of categories, descriptions of typical behaviour, the range and vari-

ations in behaviour and so on. In the case of quantitative analyses, there are known rules for conducting the analyses, and to some extent for interpreting the results obtained. For inferential statistics, the rules of both analysis and interpretation are tightly and explicitly defined. For descriptive statistics, the rules of interpretation are wider, that is, more variation is permitted by the set of conventions.

Thus there is some difference in the measure of formal control gained by using either of the two procedures. But both afford more control than numerical analyses of qualitative data or verbal analysis. Verbal analysis, in comparison with quantitative analysis, offers a greater range of interpretation since rarely is there a definitive, explicit procedure for controlling the analysis of the information collected. Case-study material, for example, may be construed in many different ways, partly because it can afford such a comprehensive account of the program under consideration. Strictly it should also be possible for data to be handled entirely through audio-visual means, excluding verbal or quantitative elements, though such treatment is rare.

In summary, formality of control in an evaluation varies with the design used and the procedures used for data collection and analyses. An evaluation study showing great formal control would use experimental design and inferential statistics, collecting data in a standardized form. An evaluation with little formal control might consist of case studies, gathering information by means that vary from person to person and with each situation, and conducting the analysis verbally. Between these two patterns lie others that combine the three elements in different ways to exhibit formal control to varying degrees. Further, it is possible for a single evaluation study to use several aspects of a single element; for example, a study may use both experimental and quasi-experimental designs, or a study may collect data in both standardized and non-standardized ways. Thus an ethnographic study of an art curriculum may use descriptive statistics for part of the analysis. An evaluation study may use a case-study approach in conjunction with an experimental or quasi-experimental design. Analyses may use both inferential and descriptive statistics. The range of possible variations of formal control is extensive. The elements which constitute the characteristic are summarized below:

A Formality of control in design:
- 1 experimental mode
  - (a) true experimental
  - (b) quasi-experimental
  - (c) non-experimental or pre-experimental

2 discursive mode
   (a) adversary proceedings
   (b) field work, systems analysis, historical analysis
   (c) criticism
   (d) impressions
3 use of both experimental and discursive modes

B Formality of data collection:
1 standardized procedures
2 idiosyncratic procedures
3 use of both standardized and idiosyncratic procedures

C Formality of data analyses:
1 quantitative analysis
   (a) inferential statistics
   (b) descriptive statistics
   (c) tabulated information
2 verbal analysis
3 use of both quantitative and verbal analyses

The elements listed above may be combined to indicate the emphasis on formality of control in a particular study. As some combinations have been indicated in the summary above, they are not repeated here.

Characteristic XI: Emphasis on Comparison

In an evaluation, sometimes comparison is of explicitly identified programs, separated out from the mass of other possible comparisons. Sometimes only one item, the program being evaluated, is clearly identified, and comparison is made with some norm. On occasion, the norm itself may not be well documented, as when comparison is with some general concept such as a 'democratic society'.

First, the program evaluated may be set against another that represents a 'no treatment' condition; or it may be set against the traditional approach in the area; or set against other new approaches. Thus an evaluator might compare success on a new physics curriculum with the traditional physics course or with other recently developed physics programs. Comparisons may also be made within a single program (Hyman & Wright, 1967); for example, instances of a program located in different places or variants at a single location. Here different versions of the same program become the object of comparison with information being gathered on their relative success.

Second, an evaluator may compare not with a second program but with a norm. He may build a comparative element into his study by using instruments that yield information for a wider population than that of the program. Statements such as, 'The median score of third-grade children in the norming population was 45.2. This is higher than third graders in the program achieved', stem from such a comparative framework. Similar use may be made of other publicly available data on demographic characteristics, voting patterns, school facilities, and so on (Weiss, 1972b). Such comparisons are with clearly defined populations, but rarely is there identification of other relevant information such as the treatment experienced by these populations.

Third, an evaluator may compare not with documented instruments and public information but with broad concepts like social justice, democracy, or the state of the economy. Here the basis for the comparison is not clearly defined. Thus, while the program being evaluated is clearly identified, the second item with which it is compared lacks clear delineation.

An evaluator may choose not to compare. His study will still have an implicit comparative element. There may be perspectives that stretch out beyond the program being examined. He may collect information on variables that have been shown in other contexts to be significant. He may even use instruments that are standardized without spelling out possible comparisons: the comparisons remain latent in the study.

Once what is to be compared is established in general terms, there remains the question of which specifics to include. Many items may be selected or only a few. An evaluator may collect identical data on two programs. When an evaluator conducts the study in this way, comparisons play an extensive role in the total evaluation.

Yet, an evaluator may collect a range of information on the program being evaluated and comparative information only for selected features. For example, an evaluator who gathers information on process as well as outcome variables for the program being evaluated may limit comparative data to outcome variables. In such a case, the role of comparison in the evaluation is much less extensive than when the information is identical for both programs. The comparative emphasis given by an evaluator may even be limited to a single element; for example, achievement scores. The way comparisons are made may also vary; one may compare directly or indirectly and in a systematic or unsystematic way. An evaluator, having collected data on home background for each of two groups, may make a direct comparison. Or he may compare the groups indirectly by means of some index or measure. For example, information on socio-economic status may

be collected by means of a scale, such as the Hollingshead Two Factor. The shift from direct to indirect comparison is important. Direct comparison lacks mediation. Nothing intervenes between features compared. Once comparison is by means of some measure, material other than that desired becomes part of it. Consequently, factors to be compared may not be easily isolated from the accompanying 'noise'.

When comparing systematically, what is to be compared is selected in advance. As part of his evaluation design, an evaluator selects program features on which he will gather comparative data. He then proceeds to collect that information and note similarities and differences. Sometimes an evaluator, while providing for possible comparison, does not identify such comparison in advance. For example, an evaluator using participant observers collects masses of information, only subsequently selecting specific items for comparison. His evaluation, while including comparison, is likely to deal with many non-comparative data. Comparisons may even be made so casually that they do not form part of the design. A development in program or evaluation may suggest a comparison that could be useful. The evaluator simply collects necessary information casually and rapidly. Here comparison is treated as incidental to the evaluation, and in an unsystematic way.

To sum up, the emphasis given to comparison in an evaluation varies with what is selected as the basic unit used, the features selected, and the approach taken. The detailed elements that have been identified are set out below:

A Unit used in the comparison:
- 1 selected groups
  - (a) among groups – the program and a 'no treatment' group and/or an alternative program (traditional or new)
  - (b) within a single program – variants at a single site or variants at several sites
- 2 one group selected and compared with a norm or concept
  - (a) built into the measuring tool
  - (b) implicit in a broad concept (e.g. social justice)
- 3 no comparison group used, the evaluator concentrates on the program alone

B Features selected for comparison:
- 1 identical information collected on the two or more programs being compared; the comparison is comprehensive
- 2 comparative information is collected for key elements addressed in the evaluation; the comparison is significant but partial

3 comparative information is limited to one or two aspects of the evaluation; the comparison is slight
4 no comparative information is gathered

C Approach taken in making the comparisons:
1 direct or indirect
(a) direct, i.e. no intervening measure
(b) mixture of direct and indirect
(c) indirect, i.e. intervening scale or measure
2 systematic or unsystematic
(a) systematic, i.e. features to be compared are selected in advance
(b) comparison provided for in the design, but the items used are selected after the data collection
(c) unsystematic, i.e. there are no planned comparisons; any comparisons are made on the basis of available information

Where the elements given in A, above, are combined with those in B and C, a reasonably clear picture of the comparative emphasis in an evaluation is given. Where the unit of comparison chosen is an alternative program or a variant of the program being evaluated, and identical data are collected for each with many features being compared directly, the evaluator has made formal and systematic comparison part of his study. Where the evaluative effort is centred on a single program, but comparisons based on normative information on key features of the program are made indirectly, the place of comparison is significant but not pervasive.

However, where an evaluator concentrates on a single program, collecting no comparative information or only doing so casually, comparison plays an unimportant part in the study. It is incidental to it.

Characteristic XII: Scope of Information Collected

The scope of information collected by an evaluator may be broad and comprehensive, represent a reasonable range, or be quite narrow and confined. The scope is comprehensive when much information on many program features is included. The information has some scope when data on several program features are included; it is narrow when the information is confined to a few program features.

Many different terms have been coined to describe evaluative information related to major aspects of a program: context, input, process,

product, antecedent, background, consequent, output, outcome, and so on. Most may be grouped under three general aspects: information relating to the background of the program prior to its introduction or at the start of the evaluation; information related to the program-in-action; and information relating to program effects. The task is to define the information that may be collected in each broad category. The list is not a definitive one but it serves to indicate the range of possibilities. Evaluations of particular programs may have items not included here; they may also include only some of those suggested.

First, there is background information. A program may go into action in a setting in which some other program has already been functioning – however well or badly. Here, background information serves to delineate the situation prior to the new program's start. Alternatively, an evaluator may be asked to evaluate a well-established program. Here too, background information is collected to form a baseline for his subsequent work. Background information may be gathered prior to a program's action phase or during it, in a retrospective manner, from records and interviews. Such information may deal with a program's target population, the context in which a program functions, and the program itself (Weiss, 1972b).

Information on the background of the target population may include sex, ethnicity, aspirations, verbal ability, intelligence, length of residence in an area, training, attitudes to the program, and support for clients to participate in it. The specific items selected will vary with the particular program. For a program whose target is multiracial classrooms, ethnicity may be an important item. For one directed at a homogeneous ethnic group, it may have little relevance. Frequently a program's target population includes program recipients and those who present the program. For example, when a new curriculum is introduced into a school program, recipients are likely to be students or children, and those who present the program, teachers and their aides. In such a case, program directors may be aiming to change not only recipients of the program but also those who directly administer it.

Information on the setting defines the relevant physical and social environment in which a program functions. It may include information on the community in which a program is based and its facilities and resources, neighbouring communities, and likely sources of support for and obstruction of the program. Information may also relate to the program itself including its rationale, statements of goals, the tactics, principles and procedures the program is built around, and its intended duration and size. An evaluator may note the resources a program uses – staff, materials, equipment, and organization.

Second, there is information related to the succession of engagements and activities that comprise the program-in-action (Stake, 1969). Such information records the program as it occurred and answers questions like these: What went on in the program? What happened? What did staff and recipients do? Such information may include how the program functions; the organization, procedures, and methods actually used; and the use made of equipment and materials. Further information on the program-in-action deals with the setting in which the program takes place. Information may deal with response to the program or with events affecting the program. For example, observations may include information on the response of those who come into contact with the program. The last set of information on program-in-action relates to the evaluator's own information-gathering techniques. On the one hand, there is the response from recipients to the techniques used – students' reaction to a test or to the presence of a constant observer. On the other, there are the possible effects of those techniques on the program itself.

Third, an evaluator may collect information on program results. Again such information may be broadly grouped into three: effects on the target population; on the environment in which the program takes place; and on the materials and staff used in the program (Hyman & Wright, 1967). There is also the possibility of collecting follow-up information on long-term effects in each of these three categories. The possible impact of a program particularly centres on the target population. Where the target population consists of individuals, information gathered may include assessment of people's attitudes, values, knowledge, and behaviour. In programs where the primary target is not the individual but an institution, a network of collaborating institutions, and even the general public, other indices of effects will be appropriate. In each case what is sought is information on the program's impact – if any. Information on the program's effects on staff may include data on their values, attitudes to the program, and knowledge of it. It may also include information on the wear and tear on staff and equipment. A program which leaves the staff exhausted at the end of eighteen months may be effective in other terms, but would require a different staffing pattern if it is to be sustained for longer periods. Information that materials supposed to last for a year come apart in six months is valuable to users.

To sum up, the scope of information collected by an evaluator may be broad and comprehensive, represent a reasonable range, or be quite narrow and confined. Information collected can include the background of the program, the program in action, and the effects of the program. The components of these sources of information are set out below:

A Information collected on the background of the program:
   1 the target population characteristics
   2 the setting – physical features and social features
   3 the program – rationale and goals, procedures, tactics, methods, equipment, materials, and staff.

B Information on the program-in-action:
   1 the functioning of the program – how it functions (organization, procedures, materials) and what happened and when (events, activities, sequences)
   2 the setting – response to the program and developments affecting the program
   3 evaluator's information-gathering techniques – reactions to them and possible effects on the program

C Information on the program effects:
   1 short-term effects
      (a) on the target population
      (b) on the setting – physical and social features
      (c) on the program staff and equipment
   2 long-term effects – on any of the above

The factors set out above may be combined to give some indication of the scope of information included in a particular evaluation study. Some examples are given below.

Where the evaluator limits data collection to selected program effects, the scope of information is narrow.

Where the evaluator collects some information on the program's rationale, on what happened when it was functioning as well as its effects on the target population and staff, the information has considerable scope.

But where an evaluator collects extensive information on the background to the program as well as its action phase and effects, the scope of information is broad and comprehensive.

Characteristic XIII: Emphasis on Interpretation

Interpretations vary considerably in the form of the argument: in logical tightness, length, and compass. As long as an interpretation is limited to strict logical inference, is tightly reasoned, and the chain of reasoning is relatively short, then alternative interpretations may be limited or even ruled out (Ennis, 1973).

When in making an interpretation an author attempts to relate specific phenomena to general qualities, the chain of reasoning

becomes longer, proceeding by other means than strict logic. Here, there may be leaps and gaps in the reasoning in some points; some issues and material may be included and others may not. The compass of the interpretation is greatly increased, as is the likelihood of alternative and even conflicting interpretations. The form of what is interpreted may also contribute to the possibility of alternative explanations. That is, where what is being interpreted is open to various constructions and, more particularly, when it is treated as a construct itself, the range of possible interpretations becomes wide, without clear-cut limits. For example, a painting is a tangible object. But what the painting is, as art, is by no means so tangible and clear (Mann, 1969). It is difficult for alternatives to be definitely ruled out. Similar difficulty is experienced with social science data representing constructs such as anxiety, intelligence, economic depression, and so on (Anderson, et al., 1975).

The key locus for interpretation in evaluation stems from the information collected. In dealing with the information he has gathered, the evaluator is likely to interpret, and present his interpretations in his reports.

In presenting a report, an evaluator may comment on the data extensively. Those comments may include broad generalities which may even be free of specific referents, thus leaving the way open for alternative interpretations. But an evaluator may also minimize any interpretation given in reports. Data may be presented with comments limited to a restatement of what lies in the data. For example, an evaluator may note that the F tests show no significant difference between students assigned to the experimental and control groups. The evaluator simply lays out the data and offers no interpretation. Alternatively, an evaluator's interpretation may be more extensive. In discussing the information collected in the study, he may offer a wide-ranging interpretation that places the information collected in the context of generalities, such as the 'technological age we live in', 'democracy', 'freedom of the individual', 'equality', and so on. By including such generalities, the evaluator has greatly increased the material relevant to the discussion and the various lines of reasoning possible. While the interpretation is still woven around the information collected, the chain of reasoning used is complex and diffuse. Conclusions reached may be plausible and reasonable, but they are not irrefutable.

In each of the instances just considered, it has been assumed that what is being interpreted is itself clear-cut. Yet, as noted above, if this is not so, alternative interpretations become possible and the role of interpretation extends to the starting point itself.

When operationalized goals, controlled procedures for data collection and analysis are used in an evaluation, the evaluator has a clear-cut starting point for any subsequent interpretation. However, as each of these features of an evaluation study becomes loosely defined, the starting point for subsequent interpretation is itself open to alternative constructions. In effect, the task of defining the starting point of the interpretation or *what* is being evaluated itself becomes part of the interpretation. When an evaluator works with goals stated in broad terms, or even more loosely defined non-referential terms, the part played by interpretation in the evaluation is extensive. Similarly, when procedures used for collecting and analysing data are only partially explicit or are left implicit, the part played by interpretation is enhanced. The effect of using non-operationalized goals and ways of handling data in an evaluation study is to make the point, from which interpretation would otherwise start, itself open to interpretation. Theoretically, an evaluator may even deal with the situation by incorporating into the study several separate constructions of the data to serve as starting points for subsequent interpretation. In practice, such an approach is too time consuming and expensive to be likely.

To sum up, there are two elements that indicate the extent to which an evaluator uses interpretation as part of his work. In the first place, there is the form of the interpretation. Second, what he interprets, or the starting point for his interpretation, may itself increase the evaluator's emphasis on interpretation. Where the evaluator's starting point is clear-cut and definite, little interpretation may be required; if the starting point is loosely defined, the evaluator may have to offer an interpretation prior to starting on the implications of the study. The characteristic is summarized below:

A Form of the interpretation:
1. the evaluator simply restates the data; the data constitute the interpretation; the interpretation is certain; alternative interpretations are excluded
2. the evaluator makes inferences from specific data; the chain of reasoning is brief; the interpretations are convincing; alternative interpretations are unlikely
3. the evaluator interprets the data in the context of broad generalities (e.g. democracy); the chain of reasoning is complex; the conclusions reached are linked to the data in general terms; the interpretations are reasonable or convincing; alternative interpretations are likely
4. the evaluator suggests alternative interpretations; the interpretations are plausible and open-ended.

B Starting point for the interpretation:
1 clear-cut
2 ambiguous
3 loosely defined

The elements in interpretation may be combined to indicate the overall interpretative quality of the study.

Where operationally defined goals and reproducible procedures for data collection and analysis are combined with an interpretation limited to restating the data, the role of interpretation in an evaluation is minimal. No latitude is allowed for alternative interpretations – although that is *not* to rule out possible disagreement with what has been done in the evaluation.

Where goals in the form of broadly stated questions and partially explicit procedures for data collection and analysis are combined with an interpretation using broad generalities and complex chains of reasoning, the role of interpretation is extensive. There is latitude for alternative interpretations, those given being reasonable not irrefutable.

Where non-referential loosely defined goals and implicit procedures for data collection and analysis are used, the role of interpretation in the evaluation study is yet further enhanced. Interpretations given may be plausible but the task of interpreting is open-ended.

5

Examples of Use of the Characteristics to Describe Evaluations

The purpose of this chapter is to illustrate how the set of characteristics developed in Chapter 3 facilitates the comparison of one evaluation with another, while indicating variations in practice among them. The chapter also serves to demonstrate how the descriptions of evaluation, developed by using the characteristics, on the one hand contain more information than a numerical rating, while on the other are still more structured than a free-ranging verbal account. Finally, the chapter illustrates the reasoning that goes into the use of the characteristics – the weighing of evidence and the difficulties encountered.

In order to demonstrate the use of the characteristics, a group of evaluation studies was selected. These were chosen to show a range of evaluative techniques and provide examples of projects with aspects which might be judged to have been unsuccessful as well as those which were successful. To avoid the influence of a single theoretician or evaluation approach, the studies selected have been done in different places and under varying conditions. A further requirement was that each study have a written account of what happened in the course of the evaluation project, not simply a report of the evaluation results or the design. This last restriction reduced the number of studies which could be used to a relatively small number. Sieber commented on the paucity of such studies: '. . . there are few specimens of live research that can be subjected to critical examination' (Sieber, 1972, p.3).

Eight studies were chosen:

1 The Neighborhood Benefit Program (NBP) (Weiss & Rein, 1969) aimed at changing existing community institutions – social agencies, schools, employment services – to make them of greater service to a range of community groups, particularly underprivileged youths. The funding agency required evaluation of the program, and insisted that progress reports had to be methodologically rigorous. The agency was also interested in the extent to which the program could be usefully transferred elsewhere. The program had

broad aims and limited standardized procedures or actions. Few programs like it had previously been tried, let alone examined.

2 The Psychiatric Inpatients (PIN) (Gove & Lubach, 1969) experiment established a pilot program to treat psychiatric inpatients at Northern State Hospital in Washington. The project involved in-service training of staff and specified procedures for treating patients with the aim of minimizing disruption to their normal lives. The evaluation study included follow-up of patients after they left the hospital.

3 The Performance Contracting Experiment (PCE) (Gramlich & Koshel, 1974) is an account of six private firms and two teacher organizations that undertook incentive contracts for the educational program at twenty public school sites. The experiment was an attempt to determine whether, in comparison with conventional practices, private companies could bring about greater improvements in the achievement levels of educationally disadvantaged elementary and junior high school students in reading and mathematics.

4 The Experimental School (ES) (Talmage, 1973) emphasized personal instruction, ungraded classes, team teaching, parental involvement, innovative instructional methods, and curricula. It was hoped that the school would act as a district centre for trying out innovations which could then be used in other schools. The school, itself, had many of the qualities of a large urban system: political pressure groups, low reading scores, behaviour problems, vandalism, and low teacher morale.

5 The Nuffield A-Level Biological Science Project (BSP) (Kelly, 1973) was set up in 1965 to develop an ecological approach to biology at Grade 12. Units were constructed on 'The Living Community', 'The Maintenance of the Individual Organism', and 'The Organism in Relation to Its Environment'. The materials were intended to give balanced and integrated coverage of major fields of biological science, both pure and applied, including human biology. Major principles, e.g. homeostasis, were built into all units, as were features of biological investigation and social, aesthetic, and humanitarian aspects.

6 The Mathematics Project (MP) (Ormell, 1973) got under way in 1969. By the autumn several packages had been developed, each representing between two and five weeks of classroom work for students. Packages included *Indices*, *Limits*, and *Quadratic Models*. Each was accompanied by a *Teacher's Guide*, a booklet containing *Hints and Answers*, and another, *Harder Problem Supplement*. Materials were

	Name	Country	Topic	Setting	Target population	Formative/ Summative	Theory allegiance	Duration of evaluation	Author
1	Neighborhood Benefit (NBP)	US	community institutions	one urban neighbourhood	social agencies	S	experimental design	3 yrs	Weiss & Rein
2	Psychiatric Inpatients (PIN)	US	mental health	one State hospital	adult psychiatric patients	S	N/A	18 months	Gove & Lubach
3	Performance Contracting (PCE)	US	math/reading learning	national (20 sites)	elementary students	S	social argument	1 yr	Gramlich & Koshel
4	Experimental School (ES)	US	ungraded ES	one school	elementary students and staff	F/S	transactional	1 yr	Talmage
5	Biological Science Project (BSP)	UK	biology curriculum	national (48 sites)	high school students	F/S	N/A	8 yrs	Kelly
6	Mathematics Project (MP)	UK	math curriculum	national	high school students	F	N/A	2 yrs	Ormell
7	Humanities Curriculum Project (HCP)	UK	inter-disciplinary curriculum	national (36-100 sites)	junior high school age students	F/S	decision-making	4 yrs	MacDonald
8	Educational Extension Agents (EEA)	US	information dissemination	three States (7 sites)	teachers, school admin., State dept personnel	F	mixed	2 yrs	Sieber

Figure 2. Summary of the Case Studies

to give students familiarity with application of mathematical models to actual, proposed, or projected non-actual situations.

7 The Humanities Curriculum Project (HCP) (MacDonald, 1973) provided stimulus, support, and materials for teaching humanities to pupils aged 14–16. The project aimed at developing an understanding of social situations, human actions, and the controversial value issues that these raise. Collections of materials were developed in such areas as *War and Society*, *The Family*, *People and Work*, and *Living Cities*. Each collection was to be used as source material for student study and teacher-led discussions. Teachers were not to present their own views on the issues and were to protect divergence of views among students.

8 The Educational Extension Agents (EEA) project (Sieber, 1972) was of a pilot program aimed at disseminating and utilizing validated information in public education. Educational extension agents functioned as channels of communication between an information retrieval centre located in the State department of education, and personnel in the public schools. They helped school personnel and those in the education department define problems, obtain the information relevant to the problem from the centre, and facilitate its use in the schools and State department. The program was set in three States with a single project director for each State and a total of seven educational extension agents.

Each case study is examined using the framework of characteristics presented in Chapter 4. For each of the characteristics, a general summary of the case studies is presented, followed by detailed accounts of two case studies in terms of the characteristic and a table summarizing all studies. In the figures, elements of a characteristic present in the account of a case study are indicated together with some comment where appropriate. A dash (–) for an element indicates that it was not present in the case-study account.

Characteristic I: Program's Stage of Development when the Evaluator Is Appointed

Summary of the Case Studies

The evaluator of the Biological Science Project (BSP) was appointed at the beginning of program planning. Few decisions affecting the evaluation were made prior to his arrival. He was able to contribute extensively to the development of the project, working with the program team on the development of objectives, design, and organization. He was also able to contribute to decisions on the form of

program used in trials, the timing of the trials, and sites and sample selected to participate in trials. The evaluator of the Mathematics Project (MP) also was appointed at the program's inception. While he took little interest in the evaluative phase until materials had been developed and were about to be tried out in schools, he too was in a position to control aspects of the program's development that affected the evaluation.

The evaluators of the Neighborhood Benefit Program (NBP), the Humanities Curriculum Project (HCP), and the Experimental School (ES) were appointed when general design of the programs had been completed, but much remained to be done in developing strategies, procedures, and so on. Hence, in these programs, the evaluator took up his or her position in a situation where program design was going on side-by-side with implementation of it in a trial-and-error form.

The Educational Extension Agents (EEA) evaluators also commenced work at the implementation stage of program development. At one site the program was implemented as a prototype; at the other two sites, in trial-and-error phase.

Finally, the evaluators of Psychiatric Inpatients (PIN) and the Performance Contracting Experiment (PCE) were both appointed when planning had been completed and programs had specific form. In each case the evaluators were dealing with a prototype of a program that could be transferred to another setting, not one in which many changes were still being made.

Detailed Illustrations: NBP and BSP

(a) The Neighborhood Benefit Program (NBP) appointed evaluators after certain basic planning decisions about the program had been made. NBP directors had settled on broad goals and a program aiming at institutional rather than individual change (Weiss & Rein, 1971a, pp.288, 290).

The program had been deliberately located in a single community whose institutions were out of touch with local residents and in which there was a large proportion of underprivileged youth. The program directors had decided on a flexible program not directed at particular neighbourhood institutions or populations. Strategies used in the program were to be developed by trial and error. Faced with the approach selected by program directors, the evaluation team of NBP had several problems. They found the broad program goals extremely difficult to operationalize (Weiss & Rein, 1971a, p.293) and simultaneous multiple approaches of the NBP impossible to separate satisfactorily in terms of effects (Weiss & Rein, 1971a, p.290).

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Planning								
1 basic research	none	–	–	none	none	none	none	none
2 needs assessment	none	–	none	informal	informal-completed	informal-completed	informal-completed	–
3 design	still being done	completed	completed	partially done	not started	started	still being done	completed
B Implementation								
1 trial and error	starting	–	–	just starting	not started 1966–67	just starting 1969–70	just starting 1968–70	just starting
2 model program	–	–	–	–	–	–	–	–
3 prototype	–	just starting	just starting single year	–	1968	–	–	–
C Adoption	–	–	–	–	1970 +	–	1970–72	just starting
D Institutionalization	–	–	–	–	–	–	–	–
OVERALL	Evaluators appointed when general design of program was completed and trial-and-error exploration of effective approaches just starting	Evaluators appointed after completion of design and at beginning of implementation of a prototype program	Experiment designed for the evaluators; later appointed just prior to implementation of prototype	Evaluator appointed after general planning completed and prior to trial-and-error implementation	Evaluator helped to dig the 'foundations' of the program	Evaluation came into action at the beginning of trials of materials	Evaluator appointed as trials of developed materials and strategies were just starting	Evaluator appointed as implementation starting in two different forms: trial-and-error and prototype

Figure 3. Characteristic I: Program's Stage of Development when the Evaluator Is Appointed

(b) In contrast, the Biological Science Project (BSP) evaluation started with the 'first activity of the project' (Kelly, 1973, p.91). The evaluator of the BSP started work so early in the project's history that he was able to use the opportunity to contribute extensively to the early planning and shaping of the project (Kelly, 1973, p.94). Thus the evaluator contributed to the definition of objectives and the outline of the scheme including content, activities, and materials. On the basis of survey information, he drafted program objectives and content outline, revising these after discussion with the development team and working parties into the *A-Level Biology: Aims and Outline Scheme* (Nuffield A-Level Biology Project, 1965). The evaluator was able to select and organize schools for trials (Kelly, 1973, pp.94, 99-101) and involve many different groups in the evaluation (Kelly, 1973, p.93). Subsequently the evaluator worked with the project during the planning phase, the initial try-outs of materials, and its adoption by schools as part of their usual curricula offerings (Kelly, 1973, p.92). The program does not appear to have gone through either a model or prototype phase; revising, inventing, and designing went on simultaneously for several years.

Characteristic II: Program's Openness to Revision

Summary of the Case Studies

Of the case study programs, two were clearly not open to revision in any respect or kind. Both the Psychiatric Inpatients program (PIN) and the Performance Contracting Experiment (PCE) were to be accepted or rejected as a whole.

In the other programs, developers, and in some cases users, were interested in making revisions. The Mathematics Project (MP) personnel were strongly committed to the rationale and organization of the program into units grouped into a particular sequence, thus possible revisions were largely limited to specific details of prepared materials. The developers of the Neighborhood Benefit Program (NBP), the Experimental School (ES), the Humanities Curriculum Project (HCP), and the Educational Extension Agents (EEA) were interested in revisions to their programs within a broad rationale and program structure. Finally, of the eight case studies, the Biological Science Project (BSP) was the most open to revision. Developers were interested in carrying out revisions to every major aspect of the program and at several levels: specific details, policies and general procedures, and also the general framework of the program.

The Performance Contracting Experiment (PCE) is illustrative of a program in which the evaluator played no part in revision. In contrast, in the Mathematics Project (MP) one of the purposes of the evaluation was to provide information for program revision.

(a) The PCE was set up by the Office of Economic Opportunity (OEO). OEO personnel established both the goals of the experiment (increase in level of achievement in reading and mathematics by academically deficient students) and the conditions of payment to firms. Neither of these basic program aspects was open to modification during the experiment, and there is no indication that OEO thought of revising them for future occasions. Certainly the evaluation was not set up to provide information of that kind. Those participating in the PCE could alter the details of their programs, both in terms of what was used and the methods involved, at any time during the year (Gramlich & Koshel, 1974, p.111 3). Some changes were brought about by exigencies facing companies: machines that could not be repaired in time, materials that were not delivered, developers' inability to maintain supply, and other unforeseen circumstances.

> In Las Vegas ... Westinghouse had to abandon its reel to reel tape recorders when it found that students learned how to erase the tapes and record obscenities. (Gramlich & Koshel, 1974, p.111 3)

By the end of the experiment:

> ... most of the programs seemed to be gravitating towards a somewhat more individualized approach than traditional classrooms, with some but not complete use of paraprofessionals and teaching machines; a limited use of tangible student incentives but fairly heavy use of the intangible incentive of free time ... (Gramlich & Koshel, 1974, p.111 4)

But such changes were not envisaged by OEO as part of continuous revisions towards the evolution of a final product. They were not part of the experiment.

(b) The Mathematics Project (MP) was in an early stage of development when evaluation started; basic planning had been done and some materials were being implemented in the schools. In the words of the project director:

> The Project is in the position of an oil company which believes that oil is to be found in a certain region, but does not know exactly where the wells should be drilled; *some* success might reasonably be expected with sites chosen at random from within the area

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Program stage								
1 planning	some	–	–	some	some	some	–	–
2 implementation	mainly	yes	yes	mainly	mainly	some	mainly	mainly
3 adoption	–	–	–	–	some	–	–	–
4 well-established	–	–	–	–	–	–	–	–
B Who revises								
1 program developers								
(a) available	yes	yes	yes	yes	yes	yes	yes	yes
(b) interested	yes	–	OEO: no	yes	yes	yes	yes	some
2 Users or adopters								
(a) available	–	–	yes	–	yes	yes	yes	–
(b) interested	–	–	–	–	–	–	yes	–
C Scope of revisions								
(a) goals	–	–	–	–	yes	–	–	some
(b) what's used	possibly	–	–	yes	yes	yes	yes	possibly
(c) what's done	yes	–	–	yes	yes	–	largely	largely
D Level of revisions								
1 specific	yes	–	–	largely	yes	largely	largely	largely
2 policies/organization	possibly	–	–	–	yes	–	possibly	some
3 general framework	–	–	–	–	possibly	–	–	–
OVERALL	Program developers open to revision of some aspects (activities and procedures)	No sign of interest in revisions of any program aspect: adopt entire or not at all	OEO: no sign of interest in revisions; firms had some in specific details	Program developers open to revisions within framework and general procedures	Many aspects of program open to revision at several different levels	Open to revision of materials in detail within organization and goals of program	Open to revision within general rationale and structure of program	Revision of some aspects in some program locations

Figure 4. Characteristic II: Program's Openness to Revision

> defined. Early material produced by a project in accordance with its curriculum hypothesis might be regarded in a similar way. (Ormell, 1973, pp.51–52)

Evaluative information was to help improve the location of the 'wells' but not to alter basic aims and structure of the program. Revisions to materials were carefully planned. For example, information from the teachers questionnaire was used to 'suggest possible changes in overall emphasis, weight, length, etc.' The pupils post-test furnished information about the extent to which 'main themes (had) stuck in the mind of the sixth former'. There were changes in emphasis in cases where this did not occur (Ormell, 1973, p.66). The scope of possible revisions in the MP was sharply defined and largely concentrated at the level of specific details rather than matters of policy or general structure of the program.

In terms of the elements of the characteristic the openness of the two programs to revision was quite different. The materials in both programs were revised. The users in the MP were interested in the changes being made; the experimenters in the PCE were not. Their prime concern was whether performance contracting worked, regardless of the materials. In the MP, the evaluation study furnished information which was used by the developers in making the revisions. In neither case were the basic aims and structure of the program open to revision.

Characteristic III: Program Uniformity from Site to Site

Summary of the Case Studies

While the case studies show considerable variation on the characteristic, none of them achieved the uniformity of, for example, a computer-assisted-instruction program. But the Psychiatric Inpatients (PIN) had definite procedures, goals, and steps, in the treatment. Placed in another context, the program would remain much the same. For the evaluator, *what* he was to evaluate was clear.

Both the Mathematics Project (MP) and Biological Science Project (BSP) offered a measure of uniformity with some variation. In each, goals and materials were uniform although variation in teaching methods was expected and accepted by program directors. Both programs monitored the use of their materials in schools but, while the MP used the information in the evaluation, the BSP used it to increase program uniformity.

The Performance Contracting Experiment (PCE) and the Edu-

cational Extension Agents (EEA) project were alike; both had some uniform framework within which program users had extensive freedom to devise activities and procedures. The constraints were, however, different. The PCE had tightly defined goals and contracts with the firms selected for the project, but permitted firms to devise their own programs. The EEA had broader goals and a core of hardware common between various sites. Again program directors were free to devise their own procedures. The three remaining programs, the Neighborhood Benefit Program (NBP), Humanities Curriculum Project (HCP), and the Experimental School (ES) put little stress on program uniformity. The ES and the NBP were both programs that were largely free to develop as their personnel decided, to adapt to particular environments (Weiss & Rein, 1971a, p.293; Talmage, 1973, p.73). *What* was to be evaluated was not easily defined nor was it so controlled and designed as to remain reasonably uniform from one context to another.

The HCP may have gone further; the project team explicitly rejected uniformity of program across contexts. The program was to be locally constructed by teachers and students who would formulate their own goals and, to a large extent, procedures, while using, as a resource, materials and teaching strategies suggested by the central development team (MacDonald, 1971; p.165).

Detailed Illustrations: NBP and EEA

(a) The account of the Neighborhood Benefit Program (NBP) suggests that it would have shown little uniformity from one site to another if it had been used in more than a single site. The NBP's administrators wanted to produce institutional change. The goals were broad and complex – open to changing emphases as program directors started work and established themselves in the area. Further, the range of institutions it was possible to include in the program was diverse: schools, job-finding agencies, playgrounds, health services, and so on (Weiss & Rein, 1971a, p.289). Those actually included would vary from one setting to another. Program directors had no core of materials or things to be systematically used in their program. Reliance was placed on staff who devised whatever enrichment or other program they thought appropriate and used whatever procedures they wished. In another neighbourhood with different institutions and people, the program would have taken another form. It was '... an unstandardized, broad aim, virtually unreplicated action-program ...' (Weiss & Rein, 1971a, p.293). Each new community into which program

Elements of the characteristic		Case studies							
		NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
I Specified									
A What's used									
1 materials:	supplied	–	–	–	–	yes	yes	yes	largely
	specified	–	–	partially	–	–	–	–	–
	devised	largely	–	partially	largely	–	–	–	–
2 staff:	supplied	–	–	–	–	–	–	–	–
	trained	–	yes	–	–	informally	–	–	yes
	untrained	largely	–	yes	yes	–	yes	yes	–
B What's done									
1 components/stages:	set	–	largely	–	–	–	yes	–	–
	selected	–	–	–	–	yes	–	largely	–
	devised	largely	–	largely	yes	–	–	some	largely
2 sequence:	set	–	largely	–	–	–	yes	–	–
	selected	–	some	–	–	yes	–	–	–
	devised	largely	–	largely	yes	–	–	largely	largely
3 procedures:	set	–	largely	–	–	–	–	–	–
	selected	–	some	–	–	some	–	partly	partly
	devised	largely	–	largely	yes	some	largely	largely	partly
C Program goals									
1 narrow:	set	–	yes	yes	–	largely	largely	–	–
	selected	–	–	–	–	–	–	–	–
	devised	–	–	–	–	–	–	–	–
2 broad/ complex:	set	–	–	–	largely	–	–	–	–
	selected	probably	–	–	some	–	–	–	largely
	devised	–	–	–	–	–	–	largely	–
II Monitored		–	possibly	possibly	–	some	noted only	not for uniformity	–
OVERALL		No standard elements, users free to emphasize different goals, devise different activities	Minor selection by users within a generally uniform approach	Uniform goals and frame work, within it considerable variation in materials and procedures	No standard elements: users free to devise own activities, develop different emphases in goals and procedures	Uniform goals and materials; some variation in use, esp. in methods. Monitoring by central staff	Uniform goals and materials; variations in method permitted: users watched but not checked	Program to be adapted to local content: users to develop own goals, activities and procedures. Stress on diversity	Common goals, equipment, training: able to develop own procedures, emphases and sequences

Figure 5. Characteristic III: Program Uniformity from Site to Site

directors moved would experience a 'different mixture of program emphases or, considering the program as a whole, a different attempt at social action' (Weiss & Rein, 1971a, p.294).

(b) The Educational Extension Agents program (EEA), while enabling program directors to develop specific procedures used, had a common core of hardware. In each State, the program centred on the use of an information retrieval system. Further the projects, although located in three widely scattered States, did share broadly conceived goals. Staff had common training sessions and additional on-the-job training from the evaluation team's field observers (Sieber, 1972, pp.7, 10). The pattern of staffing for each site was similar, including as it did a State program director, an information retrieval centre with staff, and one or more educational extension agents. Yet there were variations. In the first place, the administration varied for each State: State A had two *counties*, State B had three *regions*, and State C had two *school districts*. Second, procedures and sequences for handling inquiries were not, to any great extent, specified in advance of program operation. Instead, program personnel were able to develop their own emphases in the goals, and evolve procedures that suited their own localities. Further, during the first year of operation, the evaluation unit encouraged State directors and field agents to try out different procedures (Sieber, 1972, pp.11–12). Each State thus developed its own variation of the project within the limits set by the hardware project goals and staffing.

Characteristic IV: Specificity of Program Objectives

Summary of the Case Studies

Of the eight case studies, the Performance Contracting Experiment (PCE) program directors specified goals in the most precise form – students' end-of-program performance on selected standardized tests. The level of performance to be accepted was carefully and explicitly set down as well as the conditions under which the performance was to be assessed.

The Biological Science (BSP), Psychiatric Inpatients (PIN), and Mathematics (MP) programs were alike in having goals that were reasonably clearly specified as complex patterns of behaviour. Program directors appear to have had criteria in mind for assessing the performance of program participants. Similarly the time frame for the goal achievement was clearly set out.

In the four remaining case studies, Educational Extension Agents

(EEA), Neighborhood Benefit (NBP), the Humanities Curriculum (HCP), and the Experimental School (ES) programs, goals were stated in broad terms and in some cases – particularly the HCP – lacked any clear referent. The NBP and HCP goals were so broad in scope as to encompass a very wide range of possibly conflicting patterns of activity and thought. In addition the ES and the HCP invited program participants to evolve their own goals, and to use the program for their own ends in possibly dissimilar ways. This emphasis was much more prominent in the HCP than in the ES program: in the HCP it was the core of the program's goals; in the ES program it occurred in two of the goals. None of these three programs had indications of the response that the program directors were seeking, and little sign of criteria by which the response would be judged successful. The EEA goals were mixed; while they were broadly defined at the project level, at the state level they were sometimes defined quite specifically, at other times comprehensively.

Detailed Illustrations: PIN and EEA

(a) The directors of the Psychiatric Inpatients (PIN) program did have specific goals they were seeking, although these may not have been defined in behavioural terms. The program was directed to treating 'as effectively and rapidly as possible' the disturbance which led to hospitalization. The patient and his family were to be assisted in formulating 'realistic plans' for the patient's subsequent life in the community. In order to facilitate the patient's return to the community, the program tried to maintain the patient's connection with it during hospitalization and to prevent his 'socialization into the institutional setting of a mental hospital'. Finally, the patient's self-esteem was to be maintained during and after the period of treatment (Gove & Lubach, 1971, p.357). Thus, while the goals did not set out specific levels of acceptable performance, they were strongly suggested. Similarly, the range of activities to which the goals referred was explicit. Program directors defined carefully the various steps and stages in the treatment given patients (Gove & Lubach, 1971, pp.357, 359). The time frame in which goals were to be accomplished, while it might vary for individual patients, was also defined (Gove & Lubach, 1971, p.362). The focus of the program was on reasonably specific goals to be realized, for the most part, in the immediate future.

(b) The Educational Extension Agents (EEA) project set a number of procedural and terminal goals: helping schools to use research results, developing exemplary information resources, assisting

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Statement of goal								
1 narrowly specified	–	–	test behaviour	–	yes	yes	–	–
2 complex behaviours or states	–	largely	–	yes	sometimes	sometimes	–	yes
3 broad abstractions	yes	–	–	–	–	–	largely	–
4 open-ended	–	–	–	–	–	–	–	–
B Area								
1 conditions and activity stated	–	–	yes	–	–	–	–	–
2 range of activities indicated	very broad	yes specified	–	broad	yes	yes	very broad	largely
3 area is indeterminate	–	–	–	–	–	–	–	–
C Specificity of acceptable performance								
1 performance and level specified	–	–	explicitly	–	–	–	–	–
2 performance indicated	–	yes	–	–	yes	yes	–	–
3 performance unstated	yes	–	–	yes	–	–	yes	largely
4 self-directed/selected	–	–	–	some	–	–	yes	–
D Time frame								
1 (a) specified	–	yes	yes	–	yes	yes	–	–
(b) indeterminate	largely	–	–	largely	–	–	–	–
(c) self-selected	–	–	–	–	–	–	largely	–
2 (a) during program	–	–	–	–	–	–	–	largely
(b) end of program	–	yes	yes	–	yes	yes	–	–
(c) long term	likely	–	–	possibly	–	yes	–	–
OVERALL	Goals broadly specified complex—probably evolve as program developed; little sign of criteria by which to recognize achievement goals	Program directors had complexes of behaviour they were seeking to alter; level of performance suggested time frame specified	Program specified form, conditions and acceptable level of performance for program participants	Broad goals, some open-ended, long-term referring to a broad range of activities/context	Goals specified at various levels; time frame specified; conscious criteria for desired performance suggested	Goals specified at various levels; time frame specified (immediate and long-term); definite criteria for desired performance suggested	Experience largely to be used by teachers and students in different way to achieve their own, not project goals	Broad diffuse goals with successful performance largely unstated; goals were successively more narrowly specified

Figure 6. Characteristic IV: Specificity of Program Objectives

school personnel in defining and analysing problems, finding solutions to problems, and offering schools a follow-up service on any problems on which the field agents had worked (Sieber, Louis, & Metzger, 1972, p.23; Sieber, 1972, p.4). Within these goals, each State director was permitted to develop his own emphasis. For example, State A argued:

> The purpose of the pilot project is to design processes to accelerate the improvement of educational practices.

But State B asserted:

> The primary objective of the proposed project is to develop and test the effectiveness of a dissemination system based upon computerized storage and retrieval of selected information and a net-work or chain of inter-personnel communication links. (Sieber, Louis, & Metzger, 1972, pp.24–25)

The State program directors added to such comprehensive goals others that were highly specific, leaving unclear the connection between the two. For example, specific goals in State A included 'to furnish a reader and printer for microfilm and microfiche' and 'to work with the State television system to produce staff development programs to transmit to local teachers' (Sieber, Louis, & Metzger, 1972, p.25). In each case, the range of activities to which the broad goals referred and acceptable levels of performance of any of them were left undefined. The evaluators summed up the EEA goals as 'multiple and diffuse' (Sieber, Louis, & Metzger, 1972, p.23). The evaluators did set out to persuade State directors of the program to put their goals into concrete terms. When initial attempts brought little response, the evaluation team composed a goals checklist and circulated it to each director. On it, goals – grouped as input, process, and outcome – were stated in reasonably specific terms; for example: 'stimulate a demand for the services of the program', and 'provide and organize space and proper conditions for storage of information resources and materials' (Sieber, Louis, & Metzger, 1972, Appendix B). It is not clear whether the questionnaire led project directors to put their goals into a more concrete form, but the administration of the checklist several times in the life of the project did enable evaluators to trace changing patterns of goal emphasis.

Characteristic V: Evaluator's Independence

Summary of the Case Studies

Of the eight case studies, evaluators of the Performance Contracting

Experiment (PCE) had the greatest independence. The Office of Economic Opportunity funded the various groups involved in the teaching aspect of this experiment and contracted separately for the testing of student performance. Evaluators were well insulated from those presenting materials in the classroom and were in no way subordinate to them. Further, although evaluators reported to program developers, the key audience for reports was within the OEO and, through that office, a wider educational community, including other likely developers of materials.

Three of the remaining case studies, Humanities Curriculum Project (HCP), Educational Extension Agents (EEA), and Psychiatric Inpatients (PIN) programs, shared a similar pattern. In each, either or both the funding and hiring of the evaluator were controlled by a group other than program developers. In each, evaluators occupied senior positions in the overall organizational hierarchy with direct access to key personnel in funding institutions and program. Evaluators, while in close contact with the program, were nevertheless separate from it; they were not part of program staff. Within this general pattern, specifics varied. EEA evaluators achieved close constant contact, and yet remoteness from the program, by dividing the evaluation unit into a section located at program sites – the field observers – and a central unit, whose direct contact with the program was limited to occasional site visits. HCP evaluators shared only some facilities with program developers.

In the Neighborhood Benefit Program (NBP) evaluators had distinctly less autonomy than did the HCP or EEA evaluation – and they seem not to have fully realized the limitations on their independence. True, the evaluation unit was not part of the program team, and was not located at program headquarters; the evaluators ran a separate unit located in the same neighbourhood as the project. Yet program personnel hired the evaluators and controlled evaluation funds. Evaluators had no direct access to the agency funding the project.

In each of the remaining studies, the Experimental School (ES) program, Biological Science Project (BSP), and Mathematics Project (MP), evaluators dealt with a single group. Program developers hired the evaluator and formed the major group to whom he or she reported. Of these three, the ES evaluator was the best insulated from demands of program personnel. However, for reporting purposes, she lacked direct access to the district's community, school board, and, at some stages, even the superintendent. The other two evaluators, unlike Talmage in ES, were a part of the program staff, each occupying a senior position in the organizational structure: Kelly as co-organizer of the

BSP, and Ormell combining the roles of program and evaluation director in one. Yet there were differences. These served to check Kelly's commitment to BSP but not Ormell's to the MP. In the MP, no steps were taken to insulate the evaluation work from program development and commitments. In the BSP a network of checks on the evaluator's commitment to the program was developed. Thus, over the eight case studies, only Ormell, in evaluating the MP, lacked insulation from the program, and had little autonomy other than that attributable to his personal qualities.

Detailed Illustrations: NBP and ES

(a) The Neighborhood Benefit Program (NBP) evaluators were appointed by the program administrators, whose funding had been given partly on the condition that their program would be evaluated (Weiss & Rein, 1971a, p.287).

The evaluation unit functioned independently of the program administration. There was little contact between the two groups. The research group saw it as their responsibility to define on their own the criteria for a successful program. Program administrators made little attempt to adapt the program to the needs of the evaluators (Weiss & Rein, 1971a, p.290). By the end of the second year, program administrators could 'not be entirely certain exactly what the research group really was studying' (Weiss & Rein, 1971a, p.291) other than that it had little to offer them. Thus, in a situation where there could have been close contact between evaluators and developers, the evaluators established and maintained distance. In doing so, they behaved independently of the program and certainly did not interfere in its functioning, but they also knew little of the forms the program was taking (Weiss & Rein, 1971a, p.293). Growing dissatisfaction with the evaluation led NBP sponsors and administrators to call in a consultant to evaluate the evaluation. As a result of the consultant's report the evaluators and administrators agreed on modifications to the evaluation study which would provide information about the program more rapidly than had been previously scheduled (Weiss & Rein, 1971a, p.292). However, one person had already left the evaluation staff. Without him the revised plan was 'unworkable'. Hence:

> The administrators of the program became increasingly unsympathetic to the research, and increasingly unwilling to sponsor the research group with their funding agency. (Weiss & Rein, 1971a, p.292)

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Who controls sanctions								
1 evaluator, who also funds program	–	–	yes	–	–	–	–	–
2 shared by several groups								
(a) funding and hiring separate from program developers	–	possibly	–	–	–	–	–	yes
(b) program developers hire evaluator but funding separate	–	–	–	–	–	–	yes	–
(c) program developers hire and fund; reports elsewhere	yes	–	–	–	–	–	–	–
3 program developers fund and hire evaluator and are sole recipients of reports	–	–	–	yes	yes	yes	–	–
B Evaluator's place in hierarchy								
1 immediate access, senior position	–	yes	yes	–	yes	yes	yes	yes
2 subordinate position	yes	–	–	yes	–	–	–	–
3 junior position	–	–	–	–	–	–	–	–
C Position of evaluation unit								
1 remote from program organization	–	–	yes	–	–	–	–	–
2 separate from program organization	yes	yes	yes	yes	–	–	yes	yes
3 part of program organization								
(a) with insulation	–	–	–	–	yes	–	–	–
(b) without insulation	–	–	–	–	–	yes	–	–
OVERALL	Evaluators had indirect access to funding agency through program directors; part of the program but little contact	Evaluators had good access to program but well insulated from it; possibly separate funding; senior positions	Evaluators in control of program personnel; little contact with program; good access to key personnel	Evaluators hired by program; subordinate position; not in close daily contact with district personnel	Evaluator part of program staff and funding; but senior position and checks to balance close contact	Evaluator part of program staff and funding; in daily contact; no insulation	Evaluator funded and housed separately from program; senior position; good access to program	Evaluator funded and housed separately from program; senior position; good access to program

Figure 7. Characteristic V: Evaluator's Independence

Next the director of the evaluation left 'with rather bad feelings all round' (Weiss & Rein, 1971a, p.293), and the evaluation ended. Thus the administrators of NBP had several sanctions over the evaluation which could be and, in this case, were exercised. They controlled the evaluation's funding (Weiss & Rein, 1971a, p.293); they could curtail the budget of the evaluation if it did not meet their expectations (Weiss & Rein, 1971a, p.292); and they could refuse to co-operate with the evaluation unit if they saw no useful purpose in the evaluation work (Weiss & Rein, 1971a, pp.291–3).

(b) The evaluator of the Experimental School (ES) had to deal with a single group combining in one locus of authority the powers of funder, hirer, program developer, and sole audience for the work of the evaluator. She was hired as an independent consultant by the superintendent of the district, acting on behalf of the school board. Funding for the evaluation came from the school district where the program for the ES had also been developed (Talmage, 1972, p.70). Talmage presented the plan for the evaluation to a committee consisting of the superintendent, his assistant, and the principal of the Experimental School. Subsequently the proposed plan was ratified by the board (Talmage, 1972, p.71). The evaluator had, in effect, negotiated for herself an agreement that she was free to follow a certain course of action; a measure of independence had been negotiated. She had a good deal of face-to-face contact with district personnel. Meetings were held to discuss the evaluation, particularly proposed instrumentation, collection and interpretation of data. Yet she was not located within the school district and so was insulated, to some extent, from program developers. While the school board would make the final decision on whether or not to continue the ES, the evaluator did not have direct access to them. Communications to the board had to go through the ES principal or an assistant to the superintendent (Talmage, 1972, pp.77–8). Using these channels, none of the evaluator's reports appeared on the board's agenda (Talmage, 1972, p.78). Hence the evaluator's position prevented successful dissemination of results to parents in the district (Talmage, 1972, pp.79–80).

Characteristic VI: Evaluator's Orientation to Value Judgments on Program Merits

Summary of the Case Studies

Of the case studies, only in Sieber's evaluation of the Educational

Extension Agents (EEA) did the evaluators consistently render value judgments on program merits. Judgments as to the strengths and weaknesses of the program were drawn up by the evaluation team, yet the standards used in judging were not laid out and little attention was paid to the judgments of others.

The evaluator of the Psychiatric Inpatients (PIN) program pointed out the strengths and weaknesses of the program under scrutiny. There was no attempt at making explicit or visible the value judgments of either those evaluating or those using the program. In similar vein, the evaluators of the Biological Science Project (BSP) collected information and raised issues for consideration by others involved with the project. Both of these evaluation studies interpreted the data collected and raised issues, leaving the final verdict to others. In three of the case studies, the evaluation was largely confined to the collection of facts. In the Mathematics Project the evaluator and the program developer were the same person. In his role of evaluator, he collected and reported data. In his role as program developer, he interpreted that data and drew conclusions about possible strengths and weaknesses of the program. Talmage's account of the Experimental School (ES) indicates that the evaluator confined attention to 'the manipulation of quantitative data in an ivory tower' (Talmage, 1972, p.81). The tasks of noting program strengths and weaknesses and passing judgment on them were left to those in the school district. The evaluators of the Performance Contracting Experiment took a similar stance. The evidence was collected, analysed, and details of the interpretation given, but no judgment on the value of performance contracting was made by the evaluators. They stayed as close as possible to the facts and collected no data from others on the program's merits.

Detailed Illustrations: HCP and EEA

(a) The Humanities Curriculum Project (HCP) evaluator saw his task as '... feeding the judgment of decision makers by promoting understanding of the decisions that bear on curricular action' (MacDonald, 1971, p.167). MacDonald characterized the work of the evaluation as 'describing the experiences of hundreds of schools ... to those who had to make judgments in this field' (MacDonald, 1973, p.89) and as a way of advancing 'understanding of the interplay of forces in curriculum innovation' (MacDonald, 1973, p.90). To this end, the evaluation collected a wealth of detailed information on a wide range of variables, including the judgments of others on the program's merits. The information generated from this process was communicated to the developers, and to principals

and teachers likely to use or using the materials. The task of assessing the significance of the information was largely left to those receiving evaluation reports. On occasion the evaluator posed issues, but the thrust of the work was towards description. The issue of assigning merit to the program was avoided by the evaluation team. Skilbeck (1974), in his review of the HCP evaluation, commented that the evaluators had described the program with a wealth of detail, but had left the task of judging program merits to others. Thus, the stance of the HCP evaluator was distinctly removed from judgment.

(b) The evaluation of the Educational Extension Agents (EEA) project differs from the other case studies in that the evaluation team deliberately altered the extent to which it offered judgments on the project in the course of the evaluation. At first, the evaluation team laid out information it had collected for project personnel to use as they wished, without making explicit judgments and recommendations of their own. The evaluation team termed these reports as 'formal, non-directive' reports (Sieber, 1972, p.14). Soon, however, the evaluation team altered its reporting towards greater informality and directiveness. During an oral presentation to program directors, an evaluation team member stated:

> There should be more visits to the field by the retrieval staff and the project directors, not to 'check up' on the field agents but to observe the situation in which they are working and learn about their problems.
>
> . . . the field agents need more clerical assistance in their offices for manual searches and for record keeping. (Sieber, 1972, pp.18–19)

Similar directness in judging program merits and recommending courses of action was shown in site visits, correspondence, and phone conversations made by members of the central evaluation team who were not located on the different project sites. The field observers, too, offered direct criticisms and recommendations, but the need to maintain close communication with the field agents whom they were observing tempered judgmental quality of the comments. When observers were 'resented for intruding themselves into the work of the field agents' (Sieber, 1972, p.29), the evaluation director ordered the observer to 'desist from intervening for the indefinite future' (Sieber, 1972, p.31).

The evaluators' final report on the program offered a wealth of information together with conclusions on the strengths and weaknesses of the program and recommendations for future pro-

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Visibility of value judgments								
1 unacknowledged	yes	mainly	yes	largely	largely	largely	largely	largely
2 casually noted	–	some	–	some	some	some	little	some
3 systematically set out	–	–	–	–	–	–	–	–
B Whose value judgments								
1 evaluator's alone	largely	largely	largely	largely	some	some	–	largely
2 several groups'	–	–	–	some	some	–	some	–
3 program developer's alone	–	–	–	–	–	some	–	–
C How information presented:								
1 only 'facts' set out	–	–	largely	largely	–	largely	largely	at first
2 strengths and weaknesses suggested	–	largely	–	–	largely	possibly	possibly	largely
3 final verdict on program merits	N/A	–	–	–	–	–	–	yes
OVERALL	No final report	Merits and demerits pointed out; no judgment given	No judgment given, but some issues raised, some pointed out; largely information presented	Information presented; left to others to judge	Evaluator gathered information; posed issues for others to consider; occasionally pointed out specific weaknesses	Information was gathered; issues perhaps raised but usually left to developers to judge merits	Evaluator gathered information; possibly raised issues; but task of judging program seen as responsibility of others	Evaluators commented on merits and demerits directly, frequently; finally, they offered overall verdict on program

Figure 8. Characteristic VI: Evaluator's Orientation to Value Judgments on Program Merits

jects of a similar kind. Thus the evaluation team did pass judgment on the program's merits. Throughout, the value judgments were those of the evaluator, and the standards used in assessing program merit were left largely implicit.

Characteristic VII: Evaluator's Orientation to Utilization of his Work

Summary of the Case Studies

The Performance Contracting (PCE) and Neighborhood Benefit (NBP) evaluations were not oriented to use of evaluative information by decision-makers. Some slight orientation of the evaluation towards use was shown by the evaluators of the Psychiatric Inpatients (PIN) and Experimental School (ES). On the basis of different elements in the characteristics, the Humanities Curriculum Project (HCP) evaluator recognized overall the importance of use, but then followed a design and procedure that avoided the issue. In effect, he too showed little orientation to use. Only the Mathematics Project (MP) evaluators and those of the Biological Science Project (BSP) and Educational Extension Agents (EEA) oriented the evaluation studies towards furnishing likely decision-makers with useful information.

Detailed Illustrations: BSP and HCP

(a) The Biological Science Project (BSP) was geared to 'decision-making throughout' its life (Kelly, 1973, p.91). 'In essence the evaluation strategy consisted of establishing the issues on which decisions had to be made . . .' Users of the evaluative information included the development team, those interested in A-level biology (e.g. biologists, school inspectors) and teachers. The latter were the 'focal group' (Kelly, 1973, p.93). Further, the evaluation strategy included the participation of likely users of the project materials in the evaluation process. Some were on working parties established to deal with particular issues; for example, objectives or the place of mathematics in the scheme. Others worked on instrumentation or helped conduct the trials and so on. Users were thus not only identified by the evaluators, but involved in the evaluation. The evaluation effort was geared to the rhythm of the development team. Kelly notes that the need to produce relevant information on time for utilization meant that on occasion the 'design and procedure of the investigations were not always as precise as possible' (Kelly, 1971, p.323). The emphasis on getting information through in time for use is shown by the initial trials

of the materials. Here, evaluation data were used to revise publications, improve the supply of materials, and clarify program objectives (Kelly, 1973, p.103).

The evaluation was not geared to collecting information relative to a single decision. Instead, many issues on which decisions had to be made were identified. The questions asked about any matter to be evaluated were first, 'What is the nature of the issue?' and second, 'What data have to be collected for evaluating it?' (Kelly, 1973, p.91). Consequently, data collected were closely related to decision-making, and evaluators collected information on several different aspects of the project including objectives (consistency, clarity), arrangements for conducting trials, emphases of the materials, achievement and background of students using materials, teachers and schools, the use of mathematics in the project materials, and so on.

Information seems to have been delivered to the appropriate user – at least on most occasions. For example, initial information on time taken by the units, various activities that classes were actually doing, and reactions of students and teachers to project materials, was reported to the development team, among others (Kelly, 1973, p.102).

It seems likely that the evaluation was reasonably successful in making its findings intelligible to some of the users it had identified. There are instances (the objectives and the place of mathematics in the project are two) when the evaluators were clearly able to successfully communicate the information gathered to groups of teachers. Further, the extent to which reports on issues were linked with discussions to obtain further feedback indicate that the evaluators worked at making their findings intelligible to others.

(b) Evaluators of the Humanities Curriculum Project (HCP) initially took the position that evaluation information must be collected for use by the decision makers – 'Evaluations may be judged by whether they get the right information to the right people at the right time' – and that there is no sense in an evaluation 'providing answers to questions no one is asking' (MacDonald, 1973, p.87). Specific decision-makers were identified: the sponsors, the LEAs, schools and examining boards, as well as the planners of the HCP. The evaluator thus stressed the importance of decision-makers other than the program developers.

The evaluators, however, found it difficult to identify what decisions the various users were likely to make.

> This ... was subsequently perceived to be unsatisfactory, prin-

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Who								
1 Identified								
(a) none	likely	likely	likely	possibly	–	–	–	–
(b) one	–	–	–	–	–	developers	–	–
(c) several	–	–	–	–	development team, teachers, etc.	–	sponsors, LEAs, development, etc.	program and policy makers
2 Involvement of decision makers								
(a) removed	yes	yes	yes	–	–	–	yes	yes
(b) planning	–	–	–	some	some	–	–	–
(c) participating	–	–	–	–	some	actively	–	–
3 Timing								
(a) evaluator's design	likely	largely	largely	largely	–	–	increasingly	yes: federal
(b) decision cycle	–	–	–	–	largely	possibly	–	yes: locally
B What								
1 Decisions identified								
(a) none	likely	–	–	–	–	–	not clearly	–
(b) single global	–	yes	yes	largely	–	–	–	–
(c) several alternatives	–	–	–	–	largely	specifically identified	–	program components
2 Information collected								
(a) general interest	largely	–	largely	largely	–	–	largely	some
(b) pertinent to decisions	–	largely	–	–	largely	specifically related	–	largely
3 Information presented								
(a) intelligible	N/A	yes	not particularly	not particularly	yes	yes	not really considered	yes: locally not federally
(b) action recommendations	–	–	–	–	yes	yes	–	yes
(c) evaluator's advocacy	–	–	–	–	–	–	–	–
OVERALL	Little attention to users or to use of information in decision making	Single global decision addressed but users not clearly identified, timing and recommendations unrelated to decisions	Little attention to users or to use of information in decision making	Single global decision addressed, users not clearly identified, timing and recommendations not related to decisions	Users decisions identified; included in evaluation study and addressed in reports	Evaluation tightly related to use: users and decisions identified; data collected related to decisions	Identifies users but then does not address them in the data collection, analysis, reporting	Particularly at local level, evaluators oriented their study towards users

Figure 9. Characteristic VII: Evaluator's Orientation to Utilization of his Work

> cipally because it assumed that these groups knew in advance what questions were appropriate – an unjustified assumption when the educational process is so little understood that the effects of intervening in it cannot be fully anticipated. (MacDonald, 1973, p.88)

Consequently the evaluators of HCP redefined their task, removing the emphasis on utilization of information.

> ... the evaluation of the Humanities Curriculum Project had to be relevant to recurring problems of educational choice, and contribute to a cumulative tradition of curriculum study. (MacDonald, 1973, p.88)

MacDonald stated that, with this restructuring of the evaluation, the issue of evaluative information arriving too late to affect decisions became unreal. It is difficult to judge the extent to which the form of the evaluation reports was adapted to facilitate communication and use of information. Informal oral reports were made to the development team. The formal evaluation reports of the HCP are vast, filled with detailed description. Likely users would have to spend considerable time digesting them.

Characteristic VIII: Attention to Program Goals

Summary of the Case Studies

Evaluators of the Psychiatric Inpatients (PIN), Performance Contracting (PCE), and Experimental School (ES) programs limited their attention to what was specified by program directors. While the evaluators of the Mathematics (MP) and Biological Science (BSP) projects largely showed a similar concentration on issues specified in program rationales etc., they possibly picked up some limited incidental information on concomitant effects of the program in action. Evaluators of the Humanities Curriculum (HCP) and Neighborhood Benefit (NBP) programs showed less absorption in the program's specifications as the sole mandate for the evaluation. They collected information on a wide range of issues of which those specified by the program were a subset. HCP evaluators were, however, more comprehensive in their treatment of this aspect than the NBP evaluators. The latter addressed only concomitant outcomes, while the former included concomitant process variables as well. Evaluators of the Educational Extension Agents (EEA) program accorded an emphasis to issues specified by program rationale and to concomitant features. They also paid some slight attention to the value of the program in a wider social context.

In none of the eight case studies did evaluators pay systematic attention to the value of the program in its social and educational context. No evaluator combined such an analysis with systematic attention to the program's concomitant features.

Detailed Illustrations: ES and MP

(a) In the evaluation of the Experimental School (ES), the data for the evaluation were gathered in terms of the objectives outlined by the principal and staff of the school in conjunction with the evaluator (Talmage, 1972, p.73). The independent variable, designated as 'flexible environment', included use of space, instructional methods, curricula, and teachers' behaviour; the dependent variable included children's cognitive and affective behaviour, and parents' attitudes to the school. The evaluator thus confined the study to the objectives of the program. There was no provision for collecting incidental information on concomitant processes and outcomes of the program, and certainly not for looking at the program's value or at alternative procedures, goals structures. The evaluator in retrospect noted:

> We should have gone beyond our stated objectives because they focused too narrowly on the defined object or characteristic under study, and rarely took into consideration the many external factors that impinged upon the objectives. (Talmage, 1972, pp.181–182)

(b) Ormell, in evaluating the Mathematics Project (MP), concentrated on the materials themselves (Ormell, 1973, pp.52–53). The curriculum project had listed six points setting out the 'curriculum hypotheses' of the project, including such statements as:

> The aim of the project is to enable the sixth former to think freely about a wide range of mathematical patterns ... Motivation is the key to developing this facility. Unless the sixth former is given a strong sense of purpose, preferably via a clear perception of how and why mathematics is useful, he is unlikely to develop the kind of inner 'freedom to think' which is being sought. (Ormell, 1973, p.49)

The evaluation was directed to seeing if the materials developed 'behaved as well under classroom conditions' as these 'curriculum hypotheses' predicted (Ormell, 1973, p.51). The thrust of the evaluation was to concentrate a 'fierce intensity of effort' on improving the quality of materials (Ormell, 1973, p.52). In so directing the study, the evaluator did not pay specific attention to concomitant aspects of the project, nor to value in a wider social context. Some incidental information on such aspects may

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
Evaluator treats								
A Program goals as								
1 incidental	largely	–	–	–	–	–	largely	–
2 integral	–	yes	yes	yes	yes	yes	–	largely
B Concomitant aspects including								
1 outcomes								
(a) incidental	–	–	–	–	possibly	possibly	–	yes
(b) integral	yes	–	–	–	–	–	yes	–
2 Processes								
(a) incidental	–	–	–	–	possibly	possibly	–	–
(b) integral	–	–	–	–	–	–	yes	yes
C Aspects outside the program as								
1 incidental	–	minimal	–	–	–	–	–	minimal
2 integral	–	–	–	–	–	–	–	–
OVERALL	Wide range of possible outcomes sought	Largely limited to program specifications	Limited to program specifications	Limited to program specifications	Largely program specification; incidental attention to concomitant features	Largely limited to program specifications; possibly some concomitant aspects	Treated program goals and concomitant features	Treated program goals and concomitant features; slight attention to value of program

Figure 10. Characteristic VIII: Attention to Program Goals

have been gathered in the space left for teachers' comments on questionnaires dealing with the materials. The account does not make clear what, if anything, was done with such information. Ormell, the chief evaluator, was project director. More than that, his dissatisfaction with present mathematics courses had led to the development of the MP (Ormell, 1974, p.47). The function of the evaluation was to improve the quality of materials which had been developed, not to examine their bases or alternatives to them.

Characteristic IX: Generalizability

Summary of the Case Studies

Of the eight case studies, three treated the programs as isolated specific instances. The evaluation of the Mathematics Project (MP) was the most closely tied to the program. Questions and discussion were specific to the program: techniques used were developed just for that program, and the sample was limited to what the team could conveniently visit. Evaluations of the Experimental School (ES) and the Biological Science Project (BSP) were similar, but included some techniques that were not entirely dependent on the program evaluated. Evaluators of Psychiatric Inpatients (PIN) and Educational Extension Agents (EEA) partly dealt with the programs as isolated instances and partly as examples of a 'type'. In each case, the way information was sought reflected the emphasis. Evaluators of Performance Contracting (PCE) and the Neighborhood Benefit (NBP) programs oriented their studies largely to dealing with programs as representing particular types of programs. In each case, the techniques and sampling supported the approach taken. The HCP evaluation, of the eight considered in this treatise, was the most widely generalizable, yet even here the whole of the evaluation was not directed to that end.

Detailed Illustrations: NBP and HCP

(a) The directors of the agency that commissioned the Neighborhood Benefit Program (NBP) made it clear that the evaluators should treat the program as a model for programs of that type. In addition, the evaluators' attention was to focus on:

> the generalizable assessments of worth of the type of program rather than the more particular assessments of what had gone right and what had gone wrong in this one instance. (Weiss & Rein, 1971a, p.288)

The sociologists who undertook the evaluation of the NBP were interested in contributing to 'the development of general knowledge'. Questions central to the evaluation were framed in these terms (Weiss & Rein, 1971a, pp.288–290). Samples included in the evaluation were systematically selected with a view to possible generalization from the specific group engaged in the NBP program to a wider population – that is not to say that the sample was suitable for such generalization (Weiss & Rein, 1971a, pp.290, 292). Similar care was shown in the selection of the core instruments. The evaluators wanted a 'high quality' of measurement; the directors of the study were determined that the measurements should have maximum validity, reliability, and discriminating power (Weiss & Rein, 1971a, p.289). Many of the wide battery of instruments chosen were well documented in terms of prior usage and interpretation. Some minor studies peripheral to the central evaluation task used less well-documented and known instruments for data collection.

> . . . observers produced a set of rather journalistic materials which managed to suggest something of the style of life in the neighborhood, but nothing more. (Weiss & Rein, 1971a, p.291)

(b) The evaluators of the Humanities Curriculum Project (HCP) did not limit their approach to the program as an example of a type. Instead they sought to develop 'an empirical rather than normative model of educational decision-making and its consequences' (MacDonald, 1971, p.167); to understand the 'forces that shape' the fate of curriculum materials in the schools (MacDonald, 1973, p.80); and to contribute to an understanding of the problems of curriculum innovation in general (MacDonald & Rudduck, 1971). Some questions to which the evaluators sought answers were program specific, but even here the evaluator set out to describe the work of the project in terms that would enable sponsors and planners of 'such ventures to weigh the value of this form of investment, and to determine more precisely the framework of support, guidance, and control' which would be appropriate (MacDonald, 1973, p.88). The evaluators set out to collect information from two overlapping samples. The larger of these, involving 100 schools, was systematically drawn to permit generalizations from the sample population to the wider population of school children. The smaller sample of 12 schools, used as the basis for detailed case studies, was drawn up partly on the grounds of possible generalizations – it overlapped the large sample – and partly on accessibility of the various sites. Thus part, perhaps the largest

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Questions treat program as								
1 isolated instance	–	some	–	yes	yes	yes	little	some
2 example of a type	yes	some	yes	–	–	–	yes	some
3 example of a research construct	–	–	–	–	–	–	yes	–
B Information								
1 sample								
(a) limited to program	–	–	–	yes	yes	yes	yes	yes
(b) inference to wider population	yes	yes	yes	–	–	–	yes	possibly
2 instrumentation								
(a) standard across wide range	largely	–	all	some	–	–	some	–
(b) standard in form or principle	minor	all	–	–	some	–	some	some
(c) home-made	–	–	–	some	some	all	some	some
C Discussion treats program as								
1 isolated instance	N/A	some	–	yes	yes	yes	some	some
2 example of a type	N/A	some	yes	–	–	–	some	some
3 instance of research construct	N/A	–	–	–	–	–	some	–
OVERALL	Questions deal with program as type; techniques fit wide range of programs	Some questions and discussion deal with program as type; form of techniques stable across programs	Questions and discussion deal with program as type; techniques fit wide range of programs	Questions and discussion limited to program; techniques mixture of standard and home-made	Questions and discussion limited to program; techniques mixture of standard in form and home-made	Questions and discussion limited to program; techniques home-made	Questions and discussion from wide perspective; techniques sample mixture	Questions and discussion partly program specific, partly deal with program as type; techniques and sample relevant beyond specific instance

Figure 11. Characteristic IX: Generalizability

part, of the sampling was designed to permit generalization. The techniques used for collecting information also combined those that were well documented with others that were program dependent. On the one hand, the core of the evaluation included a battery of 21 objective measures selected from published tests and scales. On the other, teachers kept semi-structured diaries.

Discussion of the information collected shows a similar range. Some is a detailed presentation of what happened in this project: a particular instance of curriculum innovation. Other reports examine the program as an example of a type that made heavy demands on those involved, embodying innovatory value positions and using unfamiliar approaches. Some treat it as a research construct (MacDonald & Rudduck, 1971, p.148).

Characteristic X: Formality of Control

Summary of the Case Studies

The evaluations of Performance Contracting Experiment (PCE) and Psychiatric Inpatients (PIN) exhibited considerable formality of control in design, data collection, and analyses. The evaluations of Neighborhood Benefit Program (NBP) and Experimental School (ES) also showed considerable formality of control, tempered only by the inclusion of minor unformalized elements. Each of these evaluations used some form of experimental design with systematic sampling and statistical analyses of data, resulting in largely quantitative reports. The remaining evaluations all exhibited less formality of control, but in different ways. Two of the studies, Humanities Curriculum Project (HCP) and Educational Extension Agents (EEA), consistently combined formal elements of design, more especially of data collection and analyses, with qualitative informal ones. In each, data were collected by standardized as well as idiosyncratic procedures, and data were handled quantitatively as well as verbally. In each, case studies were used for the evaluation. In the Biological Science Project (BSP) evaluation, there was attention to sampling, data collection was partially standardized, and analysis had quantitative elements as well as a strong verbal component.

Of these eight evaluations, the Mathematics Project (MP) exhibited the least formality of control. It differed from BSP in that there was no attempt to sample systematically. While data collection combined some partially standardized procedures with informal impressions, the analysis subordinated the quantitative aspects to the verbal qualitative approach.

(a) The evaluators of the Psychiatric Inpatients Program (PIN) were unable to randomly assign patients to a control and experimental group because of the nature of the program, so they elected to use as the experimental group all patients entering the hospital during six months of the program's operation and

> ... to select as a control group all patients who had entered the hospital during the year prior to the establishment of the pilot program from the same county from which the program drew its patients. (Gove & Lubach, 1969, p.362)

The two groups were compared on several variables: diagnosis, history of past hospitalization, age, education, income, location and quality of residence, general index of socio-economic status, and sex. Statistically significant differences between the experimental and control groups were found on only two variables: sex and position of family responsibility.

> ... the data was (sic) systematically checked across all variables to determine whether the differences between the two groups were dependent upon these two variables. (Gove & Lubach, 1969, p.363)

Efforts were made to standardize the data collection as far as possible. Much data came from hospital records. Where follow-up interviews were used, participants were asked a standard set of questions and their responses were coded on the spot. Interviews for those in the experimental program took place, on average, 15 months after hospitalization; those for the control group took place 13 months after. In order to 'ensure uniformity' on variables possibly affected by this discrepancy, all such variables were coded at 'exactly one year following the patient's entrance to the hospital' (Gove & Lubach, 1969, p.363). Thus, the evaluators of PIN emphasized formality of control. Results were likely to be reproducible by an independent team and the conclusions to have a fairly solid defence against alternative hypotheses.

(b) The evaluators of the Educational Extension Agents (EEA) concentrated on the study of day-to-day program operations. The overall design of the study did not fall within the province of experimentation; rather it was a sociological examination of an innovative program in a case-study format at three sites (Sieber, 1972, p.6). In many cases, standardized procedures were used for collecting data on participant training needs, descriptions of target population and clients of the project, analysis of official records,

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Design								
1 experimental mode								
(a) true experiment	–	–	–	–	–	–	–	–
(b) quasi-experimental	largely	yes	yes	largely	–	–	–	–
(c) pre-experimental	–	–	–	–	yes	yes	yes	yes
2 discursive mode								
(a) adversary	–	–	–	–	–	–	–	–
(b) field work; systems analysis; historical	minimal	–	–	–	minimal	minimal	much	much
(c) criticism	–	–	–	–	–	–	–	–
(d) impressions	–	–	–	little	some	some	some	some
B Data collection								
1 standardized	yes	yes	yes	yes	–	–	yes	yes
2 idiosyncratic	slight	–	–	slight	yes	yes	yes	yes
C Data analysis								
1 quantitative								
(a) inferential	possibly	largely	largely	some	–	–	yes	yes
(b) descriptive	possibly	–	–	some	some	possibly	yes	yes
(c) tabulated	possibly	–	–	–	some	largely	yes	yes
2 verbal analysis	possibly	–	–	–	yes	yes	yes	yes
OVERALL	Tightly controlled experimental mode with systematic sampling, quantitative analysis	Attempted tightly controlled design; high degree of formal control of design, data collection and analysis	Attempted tight control: formal design, standardized data collection procedures, quantitative analysis and systematic sampling	Largely formal design: systematic sampling, standardized data collection and largely quantitative procedures; some informal anecdotal information collected	Data collection partially standardized; but quantitative element subordinated to other aspects of the evaluation	No attention to systematic sampling; some standardized procedures, some tabulation of data but qualitative analysis of data significant	Combined formal procedures with discursive ones: some reports quantitative, others largely qualitative case studies	Combination of formal and informal procedures: attempts made to systematize data and quantify them prior to interpretation

Figure 12. Characteristic X: Formality of Control

and program goals. Other techniques were partially standardized; there were semistructured interviews with project personnel. Other data were unstandardized in form. This was particularly so in the case of weekly logs of activities kept by project directors, and the voluminous reports, tapes, and discussions that field observers sent about the field agents at work. Once coded, many of these data were tabulated and analysed quantitatively. Data analyses for the evaluation in effect combined inferential and descriptive statistical procedures with extensive use of verbal analysis. Categories for coding verbal data were developed, tried out, modified, until the evaluators found them effective. In developing categories, the evaluators were attempting to make public and explicit their major analytical constructs. (See Sieber, Louis, & Metzger, 1972, Appendix D.) The central evaluation team also attempted to control idiosyncratic interpretation of data in other ways. Those working at central office made a practice of freely criticizing each other's interpretations. Sieber described such critical discussion as 'social system of inter-person checks and balances' useful in maintaining 'circumspection of judgement' (Sieber, 1972, p.41). Further, interpretations arising from either quantitative or verbal data were checked against the other (Sieber, Louis, & Metzger, 1972, p.17). The verbal analysis of data was systematic and carefully reasoned. Evaluation reports were largely verbal with quantitative tabulations and statistical analyses playing a generally subordinate role – or being relegated to the appendixes. Thus the evaluation team attempted to use qualitative procedures in a systematic and formal way.

Characteristic XI: Emphasis on Comparison

Summary of the Case Studies

In four evaluations the emphasis on the approach to comparison was very similar. Evaluators of Neighborhood Benefit Program (NBP), Psychiatric Inpatients (PIN), Performance Contracting (PCE), and Experimental School (ES) selected specific comparison groups in addition to those participating in the program, collected identical information for all groups, and focused reports on differences and similarities. Comparison was thus systematic, extensive, and integral to those four evaluation studies. The emphasis on comparison in the Educational Extension Agents (EEA) evaluation took a different form; variants of the program in three States were compared. While reports

compared these program variants, much space was also devoted to items specific to a single State, especially in the formative stages of the evaluation. The Humanities Curriculum Project (HCP) evaluators also made use of comparison. They selected a small sample to permit intensive study of program variants. In the large sample, some key elements were treated comparatively through the use of norms. However, in the total evaluation, comparisons of any kind were only a minor part.

Evaluators of the Biological Science Project (BSP) showed an even greater concentration than HCP evaluators on the program being evaluated. In the BSP evaluation no specific comparison groups were included; comparison was limited to one or two aspects of the evaluation and made largely indirectly, through norms. Finally Ormell, in evaluating the Mathematics Project (MP), made no use of comparisons.

Detailed Illustrations: ES and EEA

(a) The evaluator of the Experimental School (ES) included two comparison groups in addition to pupils attending the new school. Group A consisted of other students who had volunteered for the school but had not been selected; Group B, of all other students in the school district of a similar grade level. Also included in the study were the parents of the three groups of students (Talmage, 1973, p.75). Thus the comparisons were between the program offered at the ES and those offered elsewhere in the school district.

As the evaluation was first designed, identical data were to be collected and reported for each group (Talmage, 1972, p.74). By the end of the first year of the evaluation, circumstances in the district had altered; the evaluator was able to gather data only for Group A. No data could be collected for Group B as the school principals 'looked upon any type of comparative data as a threat' (Talmage, 1972, p.79). In the main, identical information was still collected for the two groups. Comparative information was collected systematically both directly (from school records) and indirectly (through tests, scales, and observation schedules). While some information in the evaluation was gathered incidentally, this was not used to make comparisons. Thus, in general terms, the evaluator of ES identified explicit comparison groups, collected much the same information on each group, and used comparisons extensively in reports. The approach was systematic and explicit.

(b) Evaluators of the Educational Extension Agents (EEA) program

Elements of the characteristic	Case studies							
	NBP	PIN	PGE	ES	BSP	MP	HCP	EEA
A Unit used								
1 selected groups								
(a) among	no treatment	trad. program	trad. program	trad. program	–	–	–	–
(b) within	–	–	yes	–	–	–	some	some
2 normative base								
(a) inbuilt in instruments	largely	–	largely	largely	some	–	some	–
(b) implicit in concepts	–	–	–	–	–	–	–	–
3 single program – variants on	–	–	–	–	–	yes	–	–
B Features selected								
1 identical information	yes	yes	yes	yes	–	–	–	largely
2 key aspects	–	–	–	–	–	–	yes	–
3 1 or 2 aspects	–	–	–	–	yes	–	–	–
4 no cp. information	–	–	–	–	–	yes	–	–
C Approach used								
1 direct or indirect								
(a) direct	–	–	–	–	–	–	–	–
(b) mixture	yes	yes	–	yes	–	–	yes	yes
(c) indirect	–	–	yes	–	yes	–	–	–
2 systematic or unsystematic								
(a) systematic	yes	yes	yes	yes	–	–	–	–
(b) mixture	–	–	–	–	yes	–	yes	yes
(c) unsystematic	–	–	–	–	–	–	–	–
OVERALL	Comparison systematic, extensive, integral to evaluation	Comparison systematic, extensive, integral to evaluation	Comparison systematic, extensive, integral to evaluation	Comparison systematic, extensive, integral to evaluation	Slight comparative emphasis, through interpretation of norms	No treatment of comparison; concentration on single program	Comparisons made on key elements through norms; much attention to program as a whole	Solid comparative component, identical information collected; various approaches

115

Figure 13. Characteristic XI: Emphasis on Comparison

did not include comparisons between the program and others taking the form of 'no treatment' or alternative programs. The latter would have been difficult to find as the EEA program had 'a virtual lack of precedent in the field of information dissemination and utilization' (Sieber, 1972, p.4). Installation of the EEA in three States did, however, afford opportunities for making comparisons within the program. Target areas in each State differed, which led the evaluators to adopt a 'comparative perspective' and to 'make comparisons between different modes of operation' adopted in each State (Sieber, Louis, & Metzger, 1972, pp.7–9; Sieber, 1972, p.6). EEA evaluators set out to collect identical information in each of the three settings and to replicate observations within the three different settings (Sieber, Louis, & Metzger, 1972, p.9). However, as with the ES evaluation, circumstances did not permit a similarly comprehensive coverage in each State. The EEA director in one State was

> ... deeply skeptical and even resentful of the evaluation from the beginning, and no amount of explanation or reassurance succeeded in overcoming this attitude. (Sieber, 1972, pp.8–9)

Consequently, the flow of information from that State was consistently less than from the other two States (Sieber, 1972, pp.20, 33–34). Reports reflected the situation.

On the whole, comparisons between States were approached systematically. Much of the information afforded a basis for direct comparisons – for example, statistics showing use of the retrieval system and the analysis of numerical data supplied by field agents on sources of requests and follow-up (Sieber, 1972, p.14). Some comparisons were made indirectly through questionnaires – for example, the goals checklist (Sieber, 1972, p.12) and the questionnaires sent to clients and educational personnel (Sieber, Louis, & Metzger, 1972, p.16). Other information was collected in a more casual manner. For example, evaluators selected certain items from the accounts of field agents for comparison between the three States. Thus, in the EEA evaluation, comparisons were limited to program variants. Further, the evaluators directed a large portion of their discussion in the final report to the delineation of the program as a whole.

Characteristic XII: Scope of Information Collected

Summary of the Case Studies

In three of the case studies, Neighborhood Benefit Program (NBP),

Psychiatric Inpatients (PIN), and Performance Contracting (PCE), the scope of information gathered was narrow. All centred on the target population, and gathered both background and outcome data. PIN evaluators supplemented this with a study of the program's procedures and rationale and with some attention to the setting in which the program functioned.

Evaluators of the Experimental School (ES), Biological Science Project (BSP), and Mathematics Project (MP), while collecting information on background and outcome as well as some information about the program-in-action, did so selectively. BSP evaluators supplemented this information with a follow-up study of the students who had completed the course.

Evaluators of Humanities Curriculum Project (HCP) and Educational Extension Agents (EEA) gathered a wide range of information on each major aspect of the program. For several aspects, information was gathered from several different sources and in great detail. Here the scope of information included in the evaluations was rich and comprehensive.

Detailed Illustrations: PCE and BSP

(a) The evaluators of the Performance Contracting Experiment (PCE) made little effort to deal with the program in a comprehensive way. There was a limited amount of information collected on the background and program effects, but no process information on the program-in-action. Procedures used for selecting firms meant that some background information was available on the programs used during the experimental year. As part of grant applications, firms submitted information on their resources and program. Firms selected tended to be using modern educational technology in the form of machines, tangible and intangible incentives for students and, in some cases, paraprofessionals and relatively low pupil/teacher ratios, which permitted more 'individualized' instruction (Gramlich & Koshel, 1974, pp.III 1, 2).

Some information on the setting in which the experiment would function was also collected by OEO through the procedures used to select school districts for participation. School districts selected had at least an 80 per cent low income population, a student population deficient in reading and mathematics, and recent valid and reliable achievement test data on student performance (Gramlich & Koshel, 1974, p.II 13). Background information on the target population was available from school records. This included test scores and teacher assessments. This was supplemented by pre-testing conducted by the Batelle Institute

at each site. The pre-test information gathered was largely limited to academic achievement, primarily in reading and mathematics. The planned attitude questionnaire to parents and students was difficult to circulate in time for the start of the experiment, had a low response rate, and so was abandoned.

Testing situations were carefully monitored and planned to prevent teaching to the test and to ensure reasonably valid responses. The testing conditions were not always satisfactory. The report from Batelle on the pre-testing notes that there were many instances of poor conditions through lack of discipline, overcrowding, high temperatures, and even one instance of fire drill in the middle of the testing period (Gramlich & Koshel, 1974, p.III 8). Where conditions were too bad, students were retested a few weeks later. The information collected on program outcomes consisted of the performance of the target group on a single test in reading and mathematics (as well as the other components included in the single test booklet). No other outcome information was available for use in reports (Gramlich & Koshel, 1974, p.II 7; p.IV 14-17; p.VI 5-6).

> ... the experiment would have undoubtedly benefited from more ways of measuring the success or failure of the various instructional programs. (Gramlich & Koshel, 1974, p.VI 8)

Apart from the monitoring of the testing situations, little attention was paid by the evaluator to the form the experiment actually took during the academic year. For their study of what occurred, Gramlich and Koshel (1974, III) had to rely primarily on telephone interviews with the project directors at each site and with the OEO Experiment directors. They concluded:

> The lack of control and documentation of the teaching methods used in the experiment make it impossible to go beyond crude statements, or to say why programs did or did not work ... (Gramlich & Koshel, 1974, p.VI 5)

(b) The evaluator of the Biological Science Project (BSP) was closely associated with the development of the program. The information collected was selective, but showed considerable breadth in scope. The evaluator conducted a survey of the facilities for teaching biology at the sixth form level prior to implementation of the BSP trials. At the same time, a questionnaire sent by mail collected information on the training and motivation of teachers using the scheme, the schools in which they were working, and the background of the students whom they were teaching (Kelly, 1973, p.100; Kelly, 1972b, pp.99-107; 1972c, pp.197-205).

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Collected on background								
1 target population	intensive	yes	little	some	some	little	some	–
2 setting	little	little	little	informally	some	–	some	–
3 program	little	some	little	some	much	much	much	much
B Collected on program-in-action								
1 functioning	–	–	–	–	some	some	intensive	intensive
2 setting	–	–	–	informally	little	little	some	some
3 information gathering techniques	–	–	–	–	–	–	–	little
C Collected on program effects								
1 short-term effects								
(a) target population	planned	yes	limited	yes	yes	yes	yes	yes
(b) setting	–	minimal	–	–	–	–	yes	yes
(c) program staff and equipment	–	minimal	–	–	some	–	yes	yes
2 long-term follow-up	–	–	–	–	little	–	–	–
OVERALL	Attention to the program background and outcomes in the largest population; nothing on program in action; scope narrow	Attention to the program background and outcomes in the target population; no attention to program in action; scope narrow	Attention to program outcomes, narrowly defined as achievement scores; no attention to program in action; scope narrow	Primarily interested in outcome data – sought for both parents and students; informal information on program in action	Collected information on selected program aspects: background, in action and outcome; selective breadth in scope of information	Collected information on selected program aspects: background, in action and outcome; selective breadth in scope of information	Collected broad and varied information on background, program in action and outcomes; scope of information broad and comprehensive	Collected broad and varied information on background, program in action and outcomes; scope of information broad and comprehensive

Figure 14. Characteristic XII: Scope of Information Collected

Information on the program in action was collected through questionnaires that teachers filled out at the end of each section of work, describing how they had used the materials and their own and the students' responses to them. This was supplemented by meetings with teachers, either in local groups or at national conferences, and observations in some schools (Kelly, 1973, p.93). Trials of the program took place during constant debate about the nature of the sixth form curriculum, developments in biological research, and the social significance of the subject. Each of these factors formed part of the program's setting and was noted informally by the evaluator without being woven into the study (Kelly, 1973, p.108).

The BSP evaluator also attended to the program's outcomes. Teachers' assessments of student practical work and individual projects were monitored between schools. Student achievement was also assessed through external examinations (Kelly, 1971, pp.315–327). Materials produced by the program were subject to content analysis by research workers in biology and teachers (Kelly, 1973, p.101).

A group of students who had completed the program was 'followed-up' in their first year in employment or in tertiary education (Kelly, 1972d, pp.259–266). Information was gathered on the extent to which they understood the objectives of the scheme and the careers that they were choosing to follow. Yet no attempt was made to assess the program's impact on the schools and teachers who had used it, although Kelly spoke of teachers as the 'focal group' of the whole operation (Kelly, 1973, p.101).

Characteristic XIII: Emphasis on Interpretation

Summary of the Case Studies

In three of the case studies the evaluators sharply limited any interpretation. The starting point for possible interpretations was clear-cut and interpretation was limited to restating data with an occasional inference drawn. Overall, interpretations given by the evaluators of the Performance Contracting Experiment (PCE), Experimental School (ES), and Psychiatric Inpatients (PIN) programs were presented as certain and unarguable.

The evaluators of the Educational Extension Agents (EEA) project interpreted the reasonably clear-cut information they had collected sometimes in terms of specifics, sometimes in terms of broad generalities. Interpretations were convincing and reasonable. A similar pattern of interpretation occurred in the Biological Science Project (BSP)

evaluation, although procedures for analysis and collection were looser than in the EEA evaluation and, on occasion, the evaluator assumed that interpretations would vary with the interpreter.

The evaluator of the Mathematics Project (MP) saw the task of interpreting the information collected as complex, and drew broad generalities and inferences from the data. Interpretations made were intended to be reasonable, but not certain or unarguable.

The evaluator of the Humanities Curriculum Project (HCP) exhibited a complex and extensive pattern of interpretation. Interpretation of data, while occasionally consisting of inferences, more usually included broad generalities and was viewed as an open-ended task. Interpretations were reasonable and plausible but not exhaustive and exclusive. Others were expected to reach their own conclusions and make their own interpretations of the data and HCP program.

Detailed Illustrations: PIN and EEA

(a) The goals addressed in the evaluation of the Psychiatric Inpatients (PIN) were clear-cut and definite. The variables under study were operationally defined. Hence, data collected provided a firm starting point for any subsequent interpretation that evaluators might do. Reports presented carefully laid-out data, confining interpretation in many instances to a restatement of the information contained in tables. For example, the evaluators commented:

> . . . there was no appreciable difference between groups in the proportion of males and females who received public assistance following hospitalization.
>
> Compared to the control group, fewer pilot program patients were dependent on home care following hospitalization . . .
>
> Among patients who were working just prior to hospitalization, a significantly greater proportion of pilot program patients, as compared to the control patients, returned to their job within a week following their return to the community. (Gove & Lubach, 1969, pp.336–7)

In presenting the final conclusion, their interpretation expands to drawing inferences from specific data:

> These findings provide no support for the hypothesis that a brief period of hospitalization will result in an increase in the proportion of the patients floundering in the community. (Gove & Lubach, 1969, p.369)

and again:

> These findings clearly challenge the contention that this type of approach is superficial because it does not correct basic intraper-

> sonal conflicts assumed to be the cause of some psychiatric disturbances. (Gove & Lubach, 1969, p.371)

In sum the interpretation given in PIN evaluation was sharply limited in scope. It largely consisted of a restatement of data presented in tables. Even the most general conclusions were closely related to specific points in the data. The chain of reasoning was short; generalities were excluded from the discussion.

(b) In the Mathematics Project (MP), the information collected and procedures used were only partially explicit and thus were themselves open to several constructions. The evaluator saw the interpretation of the data collected as an important part of the evaluation.

> One may regard evaluation ... as a continuing process; consisting of (i) a flow of relevant feedback information from the classroom to the project team, and (ii) the interpretation of this information. (Ormell, 1973, p.51)

Interpretation was necessary partly because evaluation data do not on their own 'generally solve' any problems.

> They (the data) may tell us when the material is not fully satisfactory, but they provide no substitute for the creative imagination needed to put that material right. (Ormell, 1973, p.67)

The first part of interpretation was confined to passing data through a number of 'filters' in order to establish its authenticity and suitability for subsequent consideration by program developers (Ormell, 1973, p.64). Only data considered to have reached a satisfactory level of each of the eight filters was used as the basis for further interpretation. 'Filters' included:

> the teacher's interpretation of the nature and purpose of the material, and his effectiveness in conveying this to the class
>
> the pupils' attitudes to the teacher and/or the school, in the social context. (Ormell, 1973, p.57)

The second stage in interpreting data concentrated on identifying weaknesses in the materials and planning remedial action. Thus the MP evaluator made extensive use of interpretation in the course of the evaluation. While closely reasoned, the chain of interpretation was lengthy and complex, including in its path hunches and intuitions as to what was the significance of the information gathered for the project. However, there is no sign of the evaluator including in his interpretation the possibility that others would interpret the materials differently as they approached them from a different perspective.

Elements of the characteristic	Case studies							
	NBP	PIN	PCE	ES	BSP	MP	HCP	EEA
A Form of interpretation								
1 restates data-certain	N/A	largely	largely	largely	–	–	–	–
2 infers from data-convincing	N/A	limited	some	some	some	some	some	some
3 uses generalities–reasonable arguments	N/A	–	–	–	some	some	some	some
4 offers multiple interpretations–plausible	N/A	–	–	–	rarely	–	some	–
B Starting point for interpretation								
1 clear-cut	largely	largely	largely	largely	some	–	–	some
2 ambiguous	–	–	–	–	some	some	some	some
3 loosely defined	–	–	–	–	–	–	some	–
OVERALL	Evaluators developed clear-cut starting point for subsequent interpretation; report never finished	Interpretations largely restated data or drew some inferences; intended to be unarguable and convincing	Interpretations largely restated data and drew some inferences; intended to be unarguable and convincing	Interpretations limited to restating data and drawing some inferences; intended to be convincing	Extensive interpretation of data using inferences and generalities including clarification of ambiguities; intended to be reasonable	Interpretation not limited to inferences; intended to include broad generalities and clarification of starting point; intended to be reasonable	Extensive interpretation including clarifying the starting point using broad generalities as well as inferences, sometimes offering multiple interpretations	Interpretations include inferences and broad generalities; intended to be convincing and reasonable

Figure 15. Characteristic XIII: Emphasis on Interpretation

6

Postscript

SUMMARY

What I have tried to do in developing the set of characteristics is to provide a framework which will further our understanding of the process of educational evaluation. At the practical level, several possibilities arise.

First, the set of characteristics facilitates a reasonably comprehensive and careful description of a particular evaluation. Structured summaries of the evaluation may be developed. These descriptions may be abbreviated, first to the form illustrated by the charts included in Chapter 5, and then to overall summary statements. By attending to each of the characteristics and their elements in turn, it is possible to develop complete and reasonably accurate summaries for an evaluation. On occasions, someone familiar with the set of characteristics may well move rapidly to a general summary of the evaluation on the characteristic, using as a model for it the examples of combined elements included in each sketch of a characteristic.

Second, the set of characteristics permits comparison of the different styles used by evaluators. Two of the case studies presented in Chapter 5 have been selected for comment. Evaluators of the Psychiatric Inpatients program (PIN) and the Performance Contracting Experiment (PCE) showed a tightly structured approach to their task. What were to be evaluated – the hypothesis tested in PCE and the treatment given in PIN – were reasonably clearly defined; these aspects of the programs were not open to revision and were fairly uniform with clearly stated goals. The evaluators had a considerable measure of independence. In each case, they chose to present information rather than judge the program merits themselves. Each paid little attention to possible users of their work. In each case, evaluator's questions attended only to matters specified by those establishing the program. The studies, although exhibiting formal control, paid little attention to comparison. Each collected information on outcome rather than program background or process. Finally, each study confined interpretation to restating data or drawing limited inferences from data.

132

In contrast, evaluators of the Biological Science and Mathematics Projects were appointed early enough in the life of the programs to be faced with few already-made constraints on the evaluation. The programs evaluated had reasonably clearly defined goals and form and were open to revision – at least in terms of the materials. Each evaluator was part of the program staff; each regarded the merit of the program in a wider social context as indisputable, improvements to be sought only on specific aspects within the programs. Each designed evaluations oriented to use with little emphasis on generalizability, formality of control and comparison. Each collected information on program processes and outcomes. Each was prepared to offer sufficient interpretation to relate the information gathered to revisions of materials. Here, the style adopted by the evaluators shows more flexibility and consequent looseness of structure than that adopted by the evaluators of PIN and PCE – but one style is not necessarily 'good' and the other 'bad'.

Third, the set of characteristics provides a ready framework for comparing theories and concepts of evaluation. In making such comparisons, not only are similarities and differences between the concepts clarified, but also the particular emphasis of a conceptualization becomes apparent.

As a brief example, a single characteristic, utilization, is examined in terms of ideas advanced by Scriven (1967), Suchman (1967), Weiss (1972b), Stake (1969), Guba and Stufflebeam (1970). Neither Suchman nor Scriven stressed use of the study. Suchman emphasized the importance of examining the objectives of the program, developing measurable criteria related to the objectives, and setting up a controlled situation to gauge the extent to which the objectives and possible side effects are achieved (Suchman, 1967, pp.27–32). Scriven viewed the evaluator's task as establishing the merit or worth of the program, not tangling with different groups of users and possible decisions stemming from the work (Scriven, 1966, pp.40, 49–55). Each of the remaining writers offered concepts of evaluation that emphasize utilization in some way. Stake recommended that an evaluator identify those likely to use the study, collect information relevant to their concerns, and process it in an intelligible form. He argued:

> It is a responsibility to gather information that will be useful to specific audiences ... It is the opportunity and obligation of each program evaluator to translate the evaluation requirements given by the State and Federal Governments into a plan that will gather useful information for the people who will read the report, particularly the local people. (Stake, 1969, p.315)

Yet there was no mention of involving likely users in the evaluation

process, nor of the evaluator recommending alternative courses of action. Weiss also defined use of the study as central to the evaluator's work:

> The purpose of evaluation research is to measure effects of a program against the goals it set out to accomplish as a means of contributing to subsequent decision-making about the program and improving future programming. (Weiss, 1972b, p.4)

She suggested that the evaluator identify those likely to use the study, as well as the uses to which the study is likely to be put. She went beyond Stake in terming lack of action recommendations in a report 'abdication' by the evaluator from an important task. She noted that an evaluator can even offer an analysis of likely consequences, good and bad, of alternative actions. But Weiss concluded that at this point the evaluator has moved into the realm of 'policy analyst and participant in decision-making process' (Weiss, 1972b, p.126). Guba and Stufflebeam (1970, p.15) stated that providing information for use in making decisions is integral to evaluation. The evaluator is to help likely users identify critical decisions, collect information relevant to those decisions, and report it in a form that indicates the benefits and drawbacks of particular lines of action. This concept of evaluation not only emphasizes use of evaluative information but makes close collaboration between users and the evaluator a necessity. Thus the characteristics facilitate the comparison of these different theories and concepts of educational evaluation. One implication of the above analysis is that an evaluator looking for guidance on possible utilization of his study would do well to turn to the latter writers.

Fourth, the characteristics facilitate comparison of concepts of evaluation with practice. Such comparisons draw attention to similarities and discrepancies between the two, and to aspects addressed in one and not the other. An issue of *Viewpoints* edited by Schwen (1972) contains four examples of evaluation based on Scriven's concept of evaluation. In general, Scriven's ideas on evaluation emphasize the importance of comparison between groups – Characteristic XI (Scriven, 1974) and the evaluator's role as judge of the program's merit – Characteristic VI (Scriven, 1967), and do not stress use of evaluative information – Characteristic VIII (Scriven, 1967, 1974). Yet none of the four examples of evaluations included formal comparison groups; none offered judgment on the merits of the program; and all of them showed a concern with the use of the information they generated in revising the programs (Sanders, 1971). More detailed examination of the case studies would point to discrepancies on other characteristics between the concept of evaluation advanced in theory and what was actually done in practice.

Fifth, the characteristics may also be used by an evaluator in designing his study. While the characteristics do not offer guidelines for conducting a study nor principles for solving problems, they do provide a rapid checklist of key aspects of the task he is undertaking. Further, the elements within a given characteristic present the range of possible alternatives an evaluator may be faced with or choose in conducting a study. For example, Characteristic II draws attention to the question of whether the program being evaluated is open to revision. The evaluator may realize that those administering the program are not interested in revising it and hence devote his energies to other purposes. The evaluator may also use two or more characteristics to set out alternative choices. Faced with a highly diverse program (Characteristic III), in order to record some of the actual forms realized in the program, he may decide to combine a formal approach in part of the evaluation with participant observation (Characteristic X). Hence, the set of characteristics may be useful to the evaluator by making him aware of the range of possible alternatives and the diverse issues that may be considered in conducting the evaluation. As indicated, the characteristics do not define what 'ought to be' the nature of each of the elements within that characteristic but the range and complexities of those elements. Most frequently the evaluator has, at best, only limited control over the aspects, or elements, within each of the characteristics. An understanding of the characteristic can, however, provide the evaluator with directions for improving the evaluation of a project on which he is, or will be, working.

Finally, those interested in conducting research into educational evaluation may use the set of characteristics to suggest inquiries into the nature of the task and work of evaluators. Two illustrations are given. Two or more characteristics may be selected and used as a basis for systematic examination of the pattern of interaction shown on them in various evaluations. In the case studies included in this treatise, two sets of interactions suggested the possibility of certain trade-offs in the approach an evaluator selects. For example, a high degree of independence (Characteristic V) may be bought at the price of high utilization (Characteristic VII). In the case studies, examined in Chapter 4, where an evaluator stressed use of the information from his study, the evaluation unit was closely integrated with the program. Where an evaluator worked less closely with program staff, his study showed less stress on utilization. To take another example, there may be a similar set of trade-offs between an evaluator who emphasizes extreme formality of control (Characteristic X) in his study and one who collects information about the program that is comprehensive in scope (Characteristic XII). In the case studies examined where an

evaluator conducted a tightly controlled study, information collected was sharply narrowed in scope. Where less formal control was part of the evaluator's approach, there was a corresponding increase in the scope of information collected in the study. While not in a form which permits direct judgment on the merits of an evaluation, the characteristics do draw attention to aspects of the evaluation that may warrant attention. Generally, the elements set out for each characteristic are on a continuum from tightly structured and well-controlled instances of that characteristic to loosely structured and diffuse instances. In an evaluation, swings from one end of the continuum to the other suggest possible weaknesses which deserve investigation. In these terms, some inconsistencies of approach occurred in the case studies examined in Chapter 4. The Neighborhood Benefit Program directors were flexible, prepared to adapt their program on the basis of evaluative information (Characteristic II). The evaluators chose not to orient their study to use (Characteristic VII), and the study was never completed. Comparison of the profile of the evaluation on the two characteristics suggests that the evaluators may have missed an opportunity. Use of a different combination of elements in the characteristic may have yielded a consistent approach. In either of the patterns of interaction noted, where a pattern held across a sufficiently large number of evaluation case studies, the functional relationship would warrant theoretical attention.

CONCLUSIONS

If the characteristics do reveal information about educational evaluation, it is reasonable to ask what kind of phenomenon they reveal. Clearly they suggest the present diversity of evaluation as an activity. In terms of these characteristics, evaluation is made up of a collection of disparate features, activities and procedures. One feature, activity, or procedure, is not derivable from another. While there are theories of research design, none deals with the possible uses of the information generated nor with ways of delivering information to a selected audience – both of which are also important features of evaluation. Similarly, even the most highly developed methodology of design does not incorporate principles for selecting the questions an evaluator may address in his work, nor guidelines for selecting information of suitable scope – both of which are characteristic of educational evaluation. The disparity of activities that together constitute evaluation and the range of forms in which each may be realized carry implications for those concerned with conceptualizing the area.

The characteristics draw attention to areas of evaluation that may

benefit from further conceptualization. This has been done for some areas, for example, on laying out the theories that underlie the use of experimental design and various statistical techniques. Yet each of these forms only a part of the range of designs and techniques of analysis used by evaluators. There is, at present, a lack of theories of comparable power for designs other than experimental, and for modes of analysis that do not draw on statistics. There is also a lack of concepts that explain how the different designs and analyses function in an evaluation. Each of these represents areas where work remains to be done.

The disparity of the characteristics and the range of alternatives that each includes make the task of establishing a standard by which to judge work done in the field enormously difficult. The possible interrelations between characteristics noted above afford a basis for delineating several styles for conducting an evaluation, for considering the extent to which a particular style is appropriate to the task in hand, and the extent to which it is consistently realized.

Finally, the disparate elements in the characteristics of educational evaluation suggest the importance to the area of an accumulated body of lore. The absence of case studies of evaluation is marked. Yet, in a field of practical endeavour like evaluation, it is the body of lore setting out what has been done in a specific context, what resulted and how others responded, that is valuable in forming a basis for principles which others may use in their own work.

> Government is made possible by the lore of politics derived from the immediate experience of the vicissitudes and tangles of legislating and administering. Institution of economic guidance and controls owes as much to unmediated experience of the marketplace as it does to formulas and theories . . . The law has systematized the accumulation of direct experience of actual cases in its machinery for the recording of cases and opinions as precedents which continuously monitor, supplement, and modify the meaning and application of its 'formal knowledge', its statutes. (Schwab, 1970, p.14)

An accumulation of case studies to which an evaluator had recourse would furnish him with examples of what others had done in a particular situation and enable him to use those principles in conducting his own studies. Others have noted the lack of case-study material in evaluation (Sieber, 1972; Weiss, 1966; 1970). The set of characteristics developed in this treatise, and their use in revealing the diversity of practice shown in the eight case studies, emphasizes the importance of an accumulated body of examples publicly available.

In the foregoing discussion, it is assumed that the set of characteristics of educational evaluation presented here does, indeed, identify key aspects of evaluation and is a reasonably comprehensive delin-

1137

eation of the field. Yet the work undertaken in developing a set of key characteristics for educational evaluation is essentially exploratory in nature. The set of characteristics developed requires further refinement and more thorough testing than was undertaken here. They require, in particular, use and examination by others before they may be regarded as an adequate account of the key aspects of educational evaluation.

This last chapter has tried to indicate, by laying out the way the characteristics were developed, possible implications and uses of the set, as well as limitations, and the potential benefit of this study to those whose work brings them into contact with evaluation.

188

Appendix

The Characteristics and Their Elements

Characteristic I: Program's Stage of Development when the Evaluator Is Appointed

A Planning:
- 1 basic research
- 2 needs assessment
- 3 invention and design

B Implementation:
- 1 trial and error
- 2 model program
- 3 prototype

C Adoption

D Institutionalization

Characteristic II: Program's Openness to Revision

A Program's stage of development:
- 1 planning
- 2 initial implementation
- 3 adoption
- 4 well established

B Who would carry out revisions:
- 1 program developers
  - (a) available
  - (b) interested
- 2 users or adopters
  - (a) available
  - (b) interested

C Scope of revisions:
- 1 every aspect of the program
  - (a) goals

139

(b) what is used, e.g. staff, equipment, materials
(c) what is done, e.g. activities, treatment, procedures, methods

2 selected aspects
(a) goals
(b) what is used
(c) what is done

3 single aspect
(a) goals
(b) what is used
(c) what is done

4 no aspect is open for revision

D Level of revisions:
1 specific or particular details
2 organization and policies
3 general framework, structure and rationale

Characteristic III: Program Uniformity from Site to Site

A What is used in the program:
1 established as
(a) materials – supplied
specified
one's own devising
(b) staff – supplied
trained
untrained
2 monitored by a central agency

B What is done in the program (the activities and treatment):
1 established as
(a) the program components or stages – set
selected
devised
(b) sequence – set
selected
devised
(c) procedures – set
selected
devised
2 monitored by a central agency

C Program goals:
1 established as
(a) narrow – set
selected
devised
(b) broad and complex – set
selected
devised
(c) no goals given
2 monitored by a central agency

Characteristic IV: Specificity of Program Objectives

A Statement of the goal:
1 narrowly specified
2 complex behaviours or states
3 broad abstractions
4 open-ended

B Area in which the goal is to be sought:
1 conditions as well as activities are stated
2 range of activities or field is indicated
3 area is indeterminate, i.e. lacks clear parameters

C Specificity of acceptable performance, i.e. what constitutes fulfilment of the goal:
1 performance and level are specified
2 performance is indicated
3 performance is largely unstated
4 performance is selected by program participants, staff, trainees, or clients

D Specificity of the time frame for achievement of the goal:
1 specified
indeterminate
self-selected
2 during the program
at the end of the program
long-term

Characteristic V: Evaluator's Independence

A Who controls the sanctions:
1 the evaluator, who also funds the program

141

2 shared by several groups, for example:
   (a) funding and hiring separate from the program developers
   (b) program developers hire the evaluator, but the funding is separate
   (c) program developers hire and fund the evaluator, but the evaluation reports are also sent elsewhere (e.g. project funding agency)
3 program developers fund and hire the evaluator and are the sole recipients of the evaluation reports

B The evaluator's place in the hierarchy:
1 immediate access to key project and funding personnel; has a senior level position
2 limited access to key project and funding personnel; has a subordinate position
3 reports to key project and funding personnel filtered through levels of the organization; has junior position

C Position of the evaluation unit:
1 remote from the program organization
2 separate from the program organization
3 part of the program organization
   (a) with insulation
   (b) without insulation

Characteristic VI: Evaluator's Orientation to Value Judgments on Program Merits

A Visibility of value judgments in the study:
1 unacknowledged
2 casually noted
3 systematically set out

B Whose value judgments are included:
1 the evaluator's alone
2 those of several groups or persons
3 the program developer's alone, or those of some other interested party

C The way value judgments are used:
1 not used – the data or 'facts' are simply presented
2 to suggest strengths and weaknesses in the program
3 to give a verdict on the merits of the program

Characteristic VII: Evaluator's Orientation to Utilization of his Work

A Who:
- 1 identification of likely users
  - (a) no specific user identified
  - (b) one major user identified
  - (c) several potential users identified
- 2 the involvement of the decision makers in the evaluation
  - (a) they are removed from the evaluation process
  - (b) they are involved in the planning of the evaluation
  - (c) they are participants in the evaluation
- 3 timing of the reports
  - (a) evaluator's design determines the report timing
  - (b) the decision cycle determines the report timing

B What:
- 1 identification of the decisions to be made
  - (a) none specifically identified
  - (b) a single global decision identified
  - (c) several alternative decisions identified
- 2 pertinence of information collected
  - (a) of general interest, not specifically related to the decisions
  - (b) pertinent to the decisions
- 3 presentation of the information
  - (a) intelligible, but not suggesting action
  - (b) in a format which suggests possible courses of action
  - (c) advocacy of recommended action

Characteristic VIII: Attention to Program Goals

The evaluator treats
A the program goals as:
- 1 incidental
- 2 integral

B the concomitant aspects including:
- 1 outcomes
  - (a) incidental
  - (b) integral
- 2 processes
  - (a) incidental
  - (b) integral

C the aspects outside the program as:
  1 incidental
  2 integral

Characteristic IX: Generalizability

A Questions treat the program as:
  1 isolated instance
  2 example of a type
  3 example of a research construct

B The way the information is sought in the evaluation:
  1 the sample
    (a) limited to the program
    (b) selected with possibility of inference to a wider population
  2 the instrumentation
    (a) well known and documented outside the bounds of the evaluation
    (b) the general approach and principle are known, the details would vary with each use of it, e.g. participant observation
    (c) the instruments are home-made, developed for and used in the evaluation study alone

C The discussion treats the program as:
  1 an isolated instance
  2 an example of a type
  3 an instance of a research construct

Characteristic X: Formality of Control

A Formality of control in design:
  1 experimental mode
    (a) true experimental
    (b) quasi-experimental
    (c) non-experimental or pre-experimental
  2 discursive mode
    (a) adversary proceedings
    (b) field work, systems analysis, historical analysis
    (c) criticism
    (d) impressions
  3 use of both experimental and discursive modes

B Formality of data collection:
  1 standardized procedures

2 idiosyncratic procedures
3 use of both standardized and idiosyncratic procedures

C Formality of data analyses:
1 quantitative analysis
(a) inferential statistics
(b) descriptive statistics
(c) tabulated information
2 verbal analysis
3 use of both quantitative and verbal analyses

Characteristic XI: Emphasis on Comparison

A Unit used in the comparison:
1 selected groups
(a) among groups – the program and a 'no treatment' group and/or an alternative program (traditional or new)
(b) within a single program – variants at a single site or variants at several sites
2 one group selected and compared with a norm or concept
(a) built into the measuring tool
(b) implicit in a broad concept (e.g. social justice)
3 no comparison group used, the evaluator concentrates on the program alone

B Features selected for comparison:
1 identical information collected on the two or more programs being compared; the comparison is comprehensive
2 comparative information is collected for key elements addressed in the evaluation; the comparison is significant but partial
3 comparative information is limited to one or two aspects of the evaluation; the comparison is slight
4 no comparative information is gathered

C Approach taken in making the comparisons:
1 direct or indirect
(a) direct, i.e. no intervening measure
(b) mixture of direct and indirect
(c) indirect, i.e. intervening scale or measure
2 systematic or unsystematic
(a) systematic, i.e. features to be compared are selected in advance
(b) comparison provided for in the design, but the items used are selected after the data collection

(c) unsystematic, i.e. there are no planned comparisons; any comparisons are made on the basis of available information

Characteristic XII: Scope of Information Collected

A Information collected on the background of the program:
1 the target population characteristics
2 the setting – physical features and social features
3 the program – rationale and goals, procedures, tactics, methods, equipment, materials, and staff

B Information on the program-in-action:
1 the functioning of the program – how it functions (organization, procedures, materials) and what happened and when (events, activities, sequences)
2 the setting – response to the program and developments affecting the program
3 evaluator's information-gathering techniques – reactions to them and possible effects on the program

C Information on the program effects:
1 short-term effects
(a) on the target population
(b) on the setting – physical and social features
(c) on the program staff and equipment
2 long-term effects – on any of the above

Characteristic XIII: Emphasis on Interpretation

A Form of the interpretation:
1 the evaluator simply restates the data; the data constitute the interpretation; the interpretation is certain; alternative interpretations are excluded
2 the evaluator makes inferences from specific data; the chain of reasoning is brief; the interpretations are convincing; alternative interpretations are unlikely
3 the evaluator interprets the data in the context of broad generalities (e.g. democracy); the chain of reasoning is complex; the conclusions reached are linked to the data in general terms; the interpretations are reasonable or convincing; alternative interpretations are likely
4 the evaluator suggests alternative interpretations; the interpretations are plausible and open-ended

B Starting point for the interpretation:
1 clear-cut
2 ambiguous
3 loosely defined

Bibliography

Abramson, D. A. Curriculum research and evaluation. *Review of Educational Research*, 1966, **36**, 388–395.

Alkin, M. C. *A framework for evaluation study*. Los Angeles: University of California, Graduate School of Education, 1969. (a)

_____. Evaluation theory development. *Evaluation Comment*, 1969, **2** (1), 1–7. (b)

_____. Behavioral objectives, specifications in evaluation: relevant or irrelevant. In *Western regional conference on testing problems, 18th. Proceedings.* Princeton, NJ: Educational Testing Service, 1969. (c) pp.1–13.

_____. Evaluating the cost-effectiveness of instructional programs. In P. A. Taylor & D. H. Cowley (Eds.), *Readings in curriculum evaluation.* Dubuque, Iowa: Wm C. Brown, 1972. pp.211–219.

_____, & Wingard, J. User oriented product evaluation. Los Angeles: University of California, Center for the Study of Evaluation, 1972.

Anderson, S. B., Ball, S., Murphy, R. T., & Associates. *Encyclopedia of educational evaluation.* San Francisco: Jossey-Bass, 1975.

Apple, M. W., Subkoviak, M. J., & Lufler, H. S., Jr (Eds.). *Educational evaluation: analysis and responsibility.* Berkeley, Ca.: McCutchan, 1974.

Association for Supervision and Curriculum Development. *Theories of instruction.* Washington, DC: 1968.

Atkin, J. M. Behavioral objectives in curriculum design: a cautionary note. *Science Teacher*, 1968, **35** (5), 27–30.

Baker, R. L. Curriculum evaluation. *Review of Educational Research*, 1969, **39**, 339–358.

Ball, S. Summative evaluation. In S. B. Anderson, et al. (Eds.), *Encyclopedia of educational evaluation.* San Francisco: Jossey-Bass, 1975. pp.406–408.

Bend, E. The impact of the social setting upon evaluative research. In American Institutes for Research, *Evaluative research: strategies and methods.* Pittsburgh: 1970. pp.109–128.

148

BLOOM, B. S. Mastery learning and its implications for curriculum development. In E. W. Eisner (Ed.), *Confronting curriculum reform*. Boston: Little Brown, 1971. pp.17–48.

——, Hastings, J. T., & Madaus, G. F. *Handbook of formative and summative evaluation of student learning*. New York: McGraw-Hill, 1971.

BRAUNER, C. J. The first probe. In R. H. P. Kraft, L. M. Smith, P. A. Pohland, C. J. Brauner, C. Gjerde, *Four evaluation examples: anthropological, economic, narrative, and portrayal*. (AERA Monograph Series on Curriculum Evaluation, 7). Chicago: Rand McNally, 1974. pp.77–98.

BRAYBROOKE, D., & LINDBLOM, C. E. *A strategy for decision: policy evaluation as a social process*. New York: Free Press, 1963.

BROOKS, M. The community action program as a setting for applied research. *Journal of Social Issues*, 1965, **21** (1), 29–40.

BUCKINGHAM, B. R. Notes on the derivation of scales in school subjects: with special application to arithmetic. In National Society for the Study of Education, *Standards and tests for the measurement of the efficiency of schools and school systems*. (15th Yearbook, Part I). Bloomington, Ill.: Public School Publishing Co., 1916.

CAIN, R. W., & LEE, E. C. A content analysis of the relationships between science and mathematics at the secondary-school level. In D. A. Payne (Ed.), *Curriculum evaluation: commentaries on purpose, process, product*. Lexington, Mass.: D. C. Heath, 1974. pp.300–307.

CAMPBELL, D. T. Reforms as experiments. *American Psychologist*, 1969, **24** (4), 409–429.

——. Qualitative knowing in action research. Kurt Lewin Award Address to the Society for the Study of Psychological Issues and the American Psychological Association, New Orleans, September 1, 1974.

——, & STANLEY, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963. pp.171–246.

CARLSON, R. V. A text pulling together the best in educational evaluation [review of B. R. Worthen & J. R. Sanders, *Educational evaluation: theory and practice*]. *Phi Delta Kappan*, 1974, **55**, 709.

CARO, F. G. (Ed.). *Readings in evaluation research*. New York: Russell Sage Foundation, 1971. (a)

——. Issues in the evaluation of social programs. *Review of Educational Research*, 1971, **41**, 87–114. (b)

CASSIRER, E. *An essay on man: an introduction to the philosophy of human culture*. New York: Bantam Books, 1970. (Reprint of 1944 ed.).

COHEN, D. Politics and research: evaluation of social action programs in education. *Review of Educational Research*, 1970, **40**, 213–238.

COLEMAN, J. S. Evaluating educational programs. *Urban Review*, 1969, **3** (4), 6–8.

COMMITTEE FOR ECONOMIC DEVELOPMENT. Research and Policy Committee. *Innovation in education: new directions for the American school: a statement of national policy* . . . New York: 1968.

COOK, D. L. Management control theory as the context for educational evaluation. *Journal of Research and Development in Education*, 1970, **3** (4), 13–26.

COOLEY, W. W., & LOHNES, P. R. Evaluative inquiry in education. MSS, 1973.

CRONBACH, L. J. Course improvement through evaluation. *Teachers College Record*, 1963, **64**, 672–683.

CUNNINGHAM, D. J. Evaluation of replicable forms of instruction: a classification of information needs in formative and summative evaluation. *AV Communication Review*, 1973, **21**, 351–367.

DRESSEL, P. L. *Evaluation in higher education.* Boston: Houghton & Mifflin, 1961.

——, & MAYHEW, L. B. *Critical thinking in social science: a handbook of suggestions for evaluation and teaching.* Dubuque, Iowa: Wm C. Brown, 1954.

——, & MAYHEW, L. B. *Handbook for theme analysis.* Dubuque, Iowa: Wm C. Brown, 1954.

——, & NELSON, C. H. *Questions and problems in science: test item folio no. 1.* Princeton, NJ: Educational Testing Service, 1956.

ECHOLS, J. P. The rise of the evaluation movement: 1920–42. Unpublished doctoral dissertation, Stanford University, 1973.

EISNER, E. W. Educational objectives – help or hindrance? *School Review*, 1967, **75**, 250–260. (a)

——. A response to my critics. *School Review*, 1967, **75**, 277–282. (b)

——. Instructional and expressive educational objectives: their formation and use in curriculum. In W. J. Popham, et al., *Instructional objectives.* (AERA Monograph Series on Curriculum Evaluation, 3). Chicago: Rand McNally, 1969. pp.1–18.

——. Emerging models for educational evaluation. *School Review*, 1972, **80**, 573–590.

ENNIS, R. H. On causality. *Educational Researcher*, 1973, **2** (6), 4–11.

ENTWISTLE, N. J., & NISBET, J. D. *Educational research in action.* London: University of London Press, 1972.

FARQUHAR, R., & MCCUAIG, D. Administration's Janus: the two faces of educational evaluation. *Education Canada*, 1972, **12**, 14–21.

FERMAN, L. Some perspectives on evaluating social welfare programs. *Annals of the American Academy of Political and Social Science*, 1969, **385**, 143–156.

150

FINN, J. D. Institutionalization of education. *Educational Technology*, 1969, **12**, 14–23.

———. Evaluation of instructional outcomes: extensions to meet current needs. *Curriculum Theory Network*, 1972, **8/9**, 96–114.

FOREHAND, G. A. The role of the evaluator in curriculum research. *Journal of Educational Measurement*, 1966, **3**, 199–204.

———. An evaluation system for curriculum innovation. *Teachers College Record*, 1971, **72**, 577–591.

FRYMIER, J. R. Curriculum assessment: problems and possibilities. In D. A. Payne (Ed.), *Curriculum evaluation*. Lexington, Mass.: D. C. Heath, 1974.

GEPHART, W. J. The Phi Delta Kappa Committee's evaluation model: one member's view. *Curriculum Theory Network*, 1972, **8/9**, 115–131.

GIDEONSE, H. D. Research development and the improvement of education. *Science*, 1968, **162**, 541–545.

GLASS, G. V. Two generations of evaluation models. Paper presented to the Nebraska Personnel and Guidance Association in Lincoln, Nebraska, September 20, 1968. In P. A. Taylor, & D. M. Cowley (Eds.), *Readings in curriculum evaluation*. Dubuque, Iowa: Wm C. Brown, 1972. pp.103–109.

———. *The growth of evaluation methodology*. Colorado: University of Colorado, 1969.

———. Design of evaluation studies. Paper presented at the Council for Exceptional Children Special Conference on Early Childhood Education, New Orleans, December 1969.

GLOCK, C. Y., et al. *Studies in the utilization of behavioral science, volume 1*. Stanford, Ca.: Institute for Communication Research, Stanford University, 1961.

GOULDNER, A. W. Explorations in applied social science. In A. W. Gouldner, & S. M. Miller (Eds.), *Applied sociology: opportunities and problems*. New York: Free Press, 1965. pp.5–22.

GOVE, W., & LUBACH, J. E. Evaluating an intensive treatment program. In F. G. Caro (Ed.), *Readings in evaluation research*. New York: Russell Sage Foundation, 1971. (a) (Reprint from Journal of Health and Social Behavior, 1969, **10**, 225–236).

GRAMLICH, E., & KOSHEL, P. Social experiments in education: the case of performance contracting. 1974.

GREENBERG, B. G. Evaluation of social programs. *Review of the International Statistical Institute Review*, 1968, **36**, 260–277.

GROBMAN, H. The place of evaluation in the biological sciences curriculum study. *Journal of Educational Measurement*, 1966, **3**, 205–212.

——. *Evaluation activities of curriculum projects: a starting point.* (AERA Series on Curriculum Evaluation, Monograph 2). Chicago: Rand McNally, 1968.

GUBA, E. Knowledge production and utilization in educational administration. In T. L. Eidell, & J. M. Kitchel (Eds.), *Knowledge production and utilization in educational administration.* Eugene: University of Oregon, Center for the Advanced Study of Educational Administration, 1968.

——. The failure of educational evaluation. *Educational Technology,* 1969, **9** (5), 29–39.

——, & CLARK, D. C. An examination of potential change roles in education. Paper read at the Symposium on Innovation in Planning School Curricula, Airlie House, Virginia, 1965.

——, & STUFFLEBEAM, D. L. *Evaluation: the process of stimulating, aiding and abetting insightful action.* (Monograph Series in Reading Education No. 1). Bloomington: Indiana University, 1970.

HAMINGSON, D. (Ed.). *Towards judgement: an edited compilation of the publications of the Evaluation Unit of the Humanities Curriculum Project, 1970–72.* (Centre for Applied Research in Education. Occasional publication no. 1). Norwich: University of East Anglia, 1973.

HAMMOND, R. L. Evaluation at the local level. (Project EPIC Report). Tucson, Arizona: 1967.

——. Context evaluation of instruction in local school districts. *Educational Technology,* 1969, **9** (1), 13–18.

HANSON, H. R. *Patterns of discovery.* London: Cambridge University Press, 1958.

HARRIS, C. Comments on Professor Wiley's paper, [Design and analysis of evaluation studies]. In M. C. Wittrock, & D. E. Wiley, (Eds.), *The evaluation of instruction: issues and problems.* New York: Holt, Rinehart, & Winston, 1970. pp.269–271.

——. (Ed.). *Problems in measuring change.* Madison: University of Wisconsin Press, 1963.

HASTINGS, J. T. Curriculum evaluation: the why of the outcomes. *Journal of Educational Measurement,* 1966, **3**, 27–32.

HERZOG, E. *Some guidelines for evaluative research.* Washington, DC.: US Department of Health, Education and Welfare, Children's Bureau, 1959.

HOUSE, E. R. (Ed.). *School evaluation: the politics and process.* Berkeley, Ca.: McCutchan, 1973.

HYMAN, H. H., & WRIGHT, C. R. Evaluating social action programs. In P. F. Lazarsfeld, W. H. Sewell, & H. L. Wilensky (Eds.), *The uses of sociology.* New York: Basic Books, 1967. pp.741–782.

JUDD, C. H. A look forward. In National Society for the Study of Education, *The measurement of educational products*. (17th Yearbook, Part II). Bloomington, Ill.: Public School Publishing Co., 1918.

KARMEL, L. J. *Measurement and evaluation in the school*. New York: Macmillan, 1970.

KELLY, E. F. *Curriculum evaluation and literary criticism: the explication of an analogy*. Syracuse: Center for Institutional Development, 1973.

KELLY, P. J. Implications of Nuffield A-level biological science. *School Science Review*, 1970, **52**, 272–285.

———. Evaluation studies of the Nuffield A-level biology trials: 1. Overall achievement of students. *Journal of Biological Education*, 1971, **5**, 315–327.

———. Evaluation studies of the Nuffield A-level biology trials: 2. Evaluation of specific objectives. *Journal of Biological Education*, 1972, **6**, 29–40. (a)

———. Evaluation studies of the Nuffield A-level biology trials: 3. Student characteristics and achievement. *Journal of Biological Education*, 1972, **6**, 99–107. (b)

———. Evaluation studies of the Nuffield A-level biology trials: 4. School characteristics and achievement. *Journal of Biological Education*, 1972, **6**, 197–205. (c)

———. Evaluation studies of the Nuffield A-level biology trials: 5. Students after the trials. *Journal of Biological Education*, 1972, **6**, 259–266. (d)

———. Nuffield A-level biological science project. In Schools Council, *Evaluation in curriculum development: twelve case studies*. (Schools Council Research Studies). London: Macmillan, 1973. pp.91–109.

———, & DOWDESWELL, W. H. Nuffield A-level biological science project. *Journal of Biological Education*, 1970, **4**, 251–260.

KOURILSKY, M. An adversary model for educational evaluation. *Evaluation Comment*, 1973, **4** (2), 3–6.

KUHN, T. *The structure of scientific revolutions*. 2nd ed. Chicago, Ill.: University of Chicago Press, 1970.

LARKINS, A. G., & SHAVER, J. P. Hard-nosed research and the evaluation of curriculum. Paper presented to the annual meeting of AERA, February 1969. Logan: Utah State University College of Education, 1969.

LAW, A. I. The impact of mandated evaluation on education. In *Western Regional Conference on Testing Problems, 20th. Proceedings*. Princeton, NJ: Educational Testing Service, 1971. pp.41–47.

LEVINE, M. Scientific method and the adversary model: some preliminary suggestions. *Evaluation Comment*, 1973, **4** (2), 1–3.

LEWY, A. The practice of curriculum evaluation. *Curriculum Theory Network*, 1973, **11**, 6–33.

LIEN, A. J. *Measurement and evaluation of learning*. Dubuque, Iowa: Wm C. Brown, 1967.

LIGHT, R. J., & SMITH, P. V. Choosing a future: strategies for designing and evaluating new programs. *Harvard Educational Review*, 1970, **40**, 1–28.

LINDVALL, C. M. The task of evaluation in curriculum development projects: a rationale and case study. *School Review*, 1966, **74**, 159–167.

LONGOOD, R., & SIMMEL, A. Organizational resistance to innovation suggested by research. Paper presented at the 57th Annual Meeting of the American Sociological Association, August 1962.

MACDONALD, B. The evaluation of the humanities curriculum project: a holistic approach. *Theory Into Practice*, 1971, **10**, 163–167.

_____. Humanities curriculum project. In Schools Council, *Evaluation in curriculum development: twelve case studies*. (Schools Council Research Studies). London: Macmillan, 1973. pp.80–90.

_____, & RUDDUCK, J. Curriculum research and development projects: barriers to success. *British Journal of Educational Psychology*, 1971, **41**, 148–154.

MACDONALD, J. B. Evaluation of evaluation. *Urban Review*, 1974, **7**, 3–14.

MACDONALD-ROSS, M. Behavioral objectives – a critical review. *Instructional Science*, 1973, **2**, 1–52.

MAGER, R. F. *Preparing instructional objectives*. Palo Alto, Ca: Fearon Press, 1962.

MANN, J. S. Curriculum criticism. *Teachers College Record*, 1969, **71**, 27–40.

MCCALL, W. A. *How to measure in education*. New York: Macmillan, 1922.

MERWIN, J. C. Historical review of changing concepts of evaluation. In National Society for the Study of Education, *Educational evaluation: new roles, new means*. (68th Yearbook, Part II). Chicago: University of Chicago Press, 1969. pp.6–25.

METFESSEL, N. S., & MICHAEL, W. B. A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. *Educational and Psychological Measurement*, 1967, **27**, 931–944.

METFESSEL, N. S., MICHAEL, W. B., & KIRSNER, D. Instrumentation of the taxonomy of educational objectives in behavioral terms. In D. A. Payne (Ed.), *Curriculum evaluation*. Lexington, Mass.: D. C. Heath, 1974. pp.81–85.

National School Public Relations Association. *The new ESEA* [Elementary and Secondary Act of 1965]. (Education USA. Special Report). Washington, DC: 1968.

Nuffield A-level Biology Project. *A-level biology: aims and outline scheme.* London: Nuffield Foundation, 1965.

Oakeshott, M. Political education. In M. Oakeshott, *Rationalism in politics and other essays.* London: Methuen, 1962. pp.111–136.

Ormell, C. P. Sixth form mathematics curriculum project. In Schools Council, *Evaluation in curriculum development: twelve case studies.* (Schools Council Research Studies). London: Macmillan, 1973. pp.47–69.

Payne, D. A. (Ed.). *Curriculum evaluation.* Lexington, Mass.: D. C. Heath, 1974.

Petrie, H. G. Theories are tested by observing the facts: or are they? In National Society for the Study of Education, *Philosophical redirection of educational research.* (71st Yearbook, Part I). Chicago: University of Chicago Press, 1972. pp.47–73.

Phi Delta Kappa National Study Committee on Evaluation. *Educational evaluation and decision making.* (Chairman: D. L. Stufflebeam). Itasca, Ill.: F. E. Peacock, 1971.

Phillips, D. C. *Theories, values, and education.* Melbourne: Melbourne University Press, 1971.

Popham, W. J. (Ed.). *Evaluation in education.* Berkeley, Ca.: McCutchan, 1974.

——, Eisner, E. W., Sullivan, H. J., & Tyler, L. *Instructional objectives.* (AERA Monograph Series on Curriculum Evaluation, 3). Chicago: Rand McNally, 1969.

Proger, B. B. Program evaluation: the model-building game. *Journal of Learning Disabilities*, 1971, **4**, 292–306

Provus, M. M. Evaluation of ongoing programs in the public school system. In National Society for the Study of Education, *Educational evaluation: new roles, new means.* (68th Yearbook, Part II). Chicago: University of Chicago Press, 1969. pp.242–283.

——. Evaluation or research, research or evaluation. *Educational Technology*, 1970, **10** (8), 50–54.

——. *Discrepancy evaluation for educational improvement and assessment.* Berkeley, Ca.: McCutchan, 1971.

——. The discrepancy model. In P. A. Taylor & D. M. Cowley (Eds.), *Readings in curriculum evaluation.* Dubuque, Iowa: Wm C. Brown, 1972. pp.117–127.

Quine, W. V. O. *Word and object.* New York: John Wiley, 1960.

Renzulli, J. S. The confessions of a frustrated evaluator. *Measurement and Evaluation in Guidance*, 1972, **5**, 298–305.

RIECKEN, H. W. Memorandum on program evaluation; [Ford Foundation, October 1953]. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education.* Boston: Allyn & Bacon, 1972. (a) pp.83–104.

RIPPEY, R. M. Can evaluation research be integral to innovation? *Curriculum Theory Network*, 1972, **8/9**, 45–58.

——. (Ed.). *Studies in transactional evaluation.* Berkeley, Ca.: McCutchan, 1973.

ROBINSON, R. *Definition.* London: Oxford University Press, 1954.

RODMAN, H., & KOLODNY, R. L. Organizational strains in the researcher-practitioner relationship. In A. W. Gouldner, & S. M. Miller (Eds.), *Applied sociology.* New York: Free Press, 1965. pp.93–113.

ROSENSHINE, B. Evaluation of classroom instruction. *Review of Educational Research*, 1970, **40** (2), 279–300.

ROSSI, P. H. Evaluating educational programs. *Urban Review*, 1969, **3**, 17–18.

——, & WILLIAMS, W. (Eds.). *Evaluating social programs: theory, practice, and politics.* New York: Seminar Press, 1972.

SANDERS, J. R. Comments on the case studies of formative evaluation: search for structure. *Viewpoints*, 1972, **48** (4), 119–132.

——, & CUNNINGHAM, D. J. A structure for formative evaluation in product development. *Review of Educational Research*, 1973, **43**, 217–236.

SARETSKY, G. The OEO PC [Performance Contracting] experiment and the John Henry effect. *Phi Delta Kappan*, 1972, **53**, 579–581.

SCHEFFLER, I. *The language of education.* Springfield, Ill.: Charles C. Thomas, 1960.

SCHOOLS COUNCIL. *Evaluation in curriculum development: twelve case studies.* (Schools Council Research Studies). London: Macmillan, 1973.

SCHULBERG, H. C., & BAKER, F. Program evaluation models and the implementation of research findings. *American Journal of Public Health*, 1968, **58** (1), 1248–1255.

SCHWAB, J. J. *The practical: a language for curriculum.* (Schools for the 70s. Auxiliary series). Washington, DC: National Education Association, Center for the Study of Instruction, 1970.

——. The practical: arts of eclectic. *School Review*, 1971, **79**, 493–542.

SCHWEN T. M. (Ed.). Four views of formative evaluation in instructional development. *Viewpoints*, 1972, **48** (4), 1–132.

SCRIVEN, M. The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation.* (AERA Monograph Series on Curriculum Evaluation, 1). Chicago: Rand McNally, 1967. pp.39–102.

——. The concept of evaluation. In M. W. Apple, M. J. Subkoviak, & H. S. Lufler (Eds.), *Educational evaluation: analysis and responsibility*. Berkeley, Ca.: McCutchan, 1974.

SIEBER, S. D. Formative evaluation: an exploration with case materials. New York: Bureau of Applied Social Research, 1972.

——, LOUIS, K. S., & METZGER, L. *The use of educational knowledge: evaluation of the pilot state dissemination program*. New York: Bureau of Applied Social Research, 1972.

SKILBECK, M. *Towards judgement*, ed. D. Hamingson: [review]. *Journal of Curriculum Studies*, 1974, **6** (1), 74–77.

SMITH, E. R., & TYLER, R. W. *Appraising and recording student progress*. New York: Harper, 1942.

SMITH, L. M., & POHLAND, P. A. Education, technology, and the rural highlands. In R. H. P. Kraft, L. M. Smith, P. A. Pohland, C. J. Brauner, C. Gjerde, *Four evaluation examples: anthropological, economic, narrative, and portrayal*. (AERA Monograph Series on Curriculum Evaluation, 7). Chicago: Rand McNally, 1974. pp.5–54.

——, & SCHUMACHER, S. Extended pilot trials of the aesthetic education program: a qualitative description, analysis and evaluation. St Ann, Mo.: Cemrel, Inc., 1972.

SNOW, R. A. Representative and quasi-representative designs for research on teaching. *Review of Educational Research*, 1974, **44**, 265–291.

STAKE, R. The countenance of educational evaluation. *Teachers College Record*, 1967, **68**, 523–540.

——. Evaluation design, instrumentation, data collection, and analysis of data. In J. L. Davis (Ed.), *Educational evaluation*. Columbus, Ohio: State Department of Public Instruction, 1969.

——. Measuring what learners learn. In *Alternative futures in American education*. (US Congress, 92. House Report 3606, Committee on Education and Labor). Washington, DC: 1972.

——, & DENNY, T. Needed concepts and techniques for utilizing more fully the potential of evaluation. In National Society for the Study of Education, *Educational evaluation: new roles, new means*. (68th Yearbook, Part II). Chicago: University of Chicago Press, 1969. pp.370–390.

STANLEY, J. C. Controlled field experiments as a model for evaluation. In P. H. Rossi & W. Williams (Eds.), *Evaluating social programs: theory, practice, and politics*. New York: Seminar Press, 1972. pp.67–71.

STENHOUSE, L. The humanities curriculum project. *Journal of Curriculum Studies*, 1968, **1** (1), 26–33.

——. Some limitations on the use of objectives in curriculum research and planning. *Paedagogica Europaea*, 1971, **6**, 74–83. (a).

——. The humanities curriculum project: the rationale. *Theory into Practice*, 1971, **10**, 154–162. (b)

STUFFLEBEAM, D. L. Evaluation as enlightenment for decision-making. In Association for Supervision and Curriculum Development, Commission on Assessment of Educational Outcomes, *Improving educational assessment, & An inventory of measures of affective behavior.* Washington, DC: 1969. pp.41–73.

SUCHMAN, E. A. *Evaluative research: principles and practices in public service and social action programs.* New York: Russell Sage Foundation, 1967.

——. Action for what? A critique of evaluative research; [paper presented at the annual meeting of the American Sociological Association, 1968]. In C. H. Weiss (Ed.), *Evaluating action programs: readings in social action and education.* Boston: Allyn & Bacon, 1972. (a) pp.52–84.

——. Evaluating educational programs. In F. G. Caro (Ed.), *Readings in evaluation research.* New York: Russell Sage Foundation, 1971. pp.43–48.

SULLIVAN, H. J. Objectives, evaluation, and improved learner achievement. In W. J. Popham, E.W. Eisner, H. J. Sullivan, L. L. Tyler, *Instructional objectives.* (AERA Monograph Series on Curriculum Evaluation, 3). Chicago: Rand McNally, 1969. pp.65–99.

TALMAGE, H. A case study of an evaluation project. In R. M. Rippey (Ed.), *Studies in transactional evaluation.* Berkeley, Ca.: McCutchan, 1973. pp.67–85.

TAWNEY, D. (Ed.). *Curriculum evaluation today: trends and implications.* (Schools Council Research Studies). London: Macmillan, 1976.

TAYLOR, P. A., & COWLEY, D. M. (Eds.). *Readings in curriculum evaluation.* Dubuque, Iowa: Wm C. Brown, 1972.

——, & MAGUIRE, T. O. A theoretical evaluation model. *Manitoba Journal of Educational Research*, 1966, **1**, 12–17.

THORNDIKE, E. L. Address from the proceedings of the Conference on Educational Measurements. *Indiana University Bulletin*, 1914, **12** (10).

——. *Human learning.* New York: D. Appleton, 1921.

THORNDIKE, R. L., & HAGEN, E. P. *Measurement and evaluation.* New York: John Wiley, 1961.

TOULMIN, S. *Foresight and understanding.* New York: Harper Torch Books, 1961.

TYLER, R. W. Measuring the ability to infer. *Educational Research Bulletin*, 1930, **9**, 475–480.

——. A generalized technique for conducting achievement tests. *Educational Research Bulletin*, 1931, **10**, 199–208.

1581

——. Techniques for evaluating behavior. *Educational Research Bulletin*, 1934, **13**, 1–11. (a)

——. *Constructing achievement tests*. Columbus, Ohio: Ohio State University, 1934. (b)

——. *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press, 1950.

——. Changing concepts of educational evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*. (AERA Monograph Series on Curriculum Evaluation, 1). Chicago: Rand McNally, 1967. pp.13–18.

——. (Ed.). *Educational evaluation: new roles, new means:* the sixty-eighth yearbook of the National Society for the Study of Education, part II. Chicago: University of Chicago Press, 1969.

United States National Advisory Council on Education Professions Development (NACEPD). *Evaluation of educational programs*. Washington, DC: US Office of Education, 1969.

Vickers, Sir G. The art of judgement. London: Chapman & Hall, 1965.

——. *Value systems and social process*. London: Tavistock, 1968.

Walberg, H. J. (Ed.) *Evaluating educational performance: a source book of methods, instruments, and examples*. Berkeley, Ca.: McCutchan, 1974.

Walker, D. F., & Schaffarzick, T. Comparing curricula. *Review of Educational Research*, 1974, **44** (1), 83–111.

Weiss, C. H. Planning an action project evaluation. In J. Shmelzer (Ed.), *Learning in action*. Washington, DC: US Government Printing Office, 1966.

——. Utilization of evaluation: towards comparative study. In *The use of social research in federal domestic programs*, 3. (US Congress. 90, 1st session. Committee on Government Operations, Research and Technical Programs Subcommittee). Washington, DC: US Government Printing Office, 1967, pp.426–432.

——. Politicization of evaluation research. *Journal of Social Issues*, 1970, **26** (4), 57–68.

——. (Ed.). *Evaluating action programs: readings in social action and education*. Boston: Allyn & Bacon, 1972. (a)

——. *Evaluation research: methods of assessing program effectiveness*. (Methods of Social Science Series). Englewood Cliffs, NJ: Prentice-Hall, 1972. (b)

——. Between the cup and the lip. *Evaluation Comment*, 1973, **4** (3), 1–6.

Weiss, R. S., & Rein, M. The evaluation of broad-aim programs: a cautionary case and a moral. In F. G. Caro (Ed.), *Readings in evaluation research*. New York: Russell Sage Foundation, 1971. (a)

pp.287–296. Reprinted from *Annals of the American Academy of Political and Social Science*, 1969, **385**.

——, & Rein, M. The evaluation of broad-aim programs: experimental design, its difficulties, and an alternative. *Administrative Science Quarterly*, 1970, **15** (1), 97–109.

Welch, W. W. Some problems in evaluating a national curriculum project. *Curriculum Theory Network*, 1972, **8/9**, 232–241.

——, & Walberg, H. J. A design for curriculum evaluation. *Science Education*, 1968, **52**, 10–16.

Westbury, I. Curriculum evaluation. *Review of Educational Research*, 1970, **40**, 239–260.

Wiley, D. E. Design and analysis of evaluation studies. In M. C. Wittrock, & D. E. Wiley (Eds.), *The evaluation of instruction: issues and problems*. New York: Holt, Rinehart & Winston, 1970. pp.259–273.

Wilhelms, F. T. (Ed.). *Evaluation as feedback and guide*. Washington, DC: Association for Supervision and Curriculum Development, 1967.

Wittrock, M. C., & Wiley, D. E. (Eds.). *The evaluation of instruction: issues and problems;* [symposium, University of California at Los Angeles, 1967]. New York: Holt, Rinehart, & Winston, 1970.

Wolf, R. L. The application of selected legal concepts to educational evaluation. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1974.

——. Making education accountable to the learner: framework for evaluation. *Journal of Research and Development in Education*, 1971, **5** (1), 37–48.

Wood, B. D. *Measurement in higher education*. New York: World Book Company, 1923.

Worthen, B. Toward a taxonomy of evaluation designs. *Educational Technology*, 1968, **8** (15), 3–9.

——, & Sanders, J. R. *Educational evaluation: theory and practice*. Worthington, Ohio: Charles A. Jones, 1973.

Wright, C., & Hyman, H. The evaluators. In P. E. Hammond (Ed.), *Sociologists at work*. New York: Basic Books, 1964. pp.121–141.

Wringe, S. Some problems raised by the Schools Council humanities project. *Journal of Curriculum Studies*, 1974, **6**, 30–43.

The author has developed a set of key characteristics to assist the understanding of the process of educational evaluation. These key characteristics

- facilitate a comprehensive and careful description of any evaluation study
- permit comparison of the different styles used by evaluators
- provide a framework for comparing theories and concepts of evaluation
- enable comparison of concepts of evaluation with practice.

The characteristics may also be used by an evaluator when designing a study. Those involved in conducting research into educational evaluation may use the set of characteristics to suggest inquiries into the nature of the task and the work of evaluators.

The author has made a wide survey of the ideas of those writing about evaluation in and outside education, and has tested the set of characteristics on actual instances of evaluation.

The book is essential reading for all who are interested in educational evaluation.

Australian Council for Educational Research

isbn 0 85563 180 5 161